2003

A comprehensive analysis of recently integrated human LINE-1 mobile elements

Jeremy Shawn Myers
Louisiana State University and Agricultural and Mechanical College

A COMPREHENSIVE ANALYSIS OF RECENTLY INTEGRATED HUMAN LINE-1 MOBILE ELEMENTS

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Biological Sciences

by
Jeremy Shawn Myers
B.S., Bucknell University, 1999
May 2003

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Mark A Batzer for his intellectual and financial contribution to this work, collaborators and the members of my dissertation committee Drs. David W. Foltz, John C. Larkin, Patrick J. DiMario, Fred M. Enright, and Mohamed A. F. Noor for their scientific advice and assistance in the preparation of my dissertation. I would also like to express my gratitude to all of my friends, the members of the Batzer Lab, and my family for their support. Lastly, I would like to thank Bethaney Vincent and her family for their friendship and encouragement.

I would also like to acknowledge the following collaborators: Hunt Udall, Tammy A. Morrish, Gail E. Kilroy, Dr. Gary D. Swergold, Dr. Jurgen Henke, Dr. Lotte Henke, and Dr. John V. Moran for their contribution to chapter two of this work; Dr. Abdel-Halim Salem and Anthony C. Otieno for their contribution to chapter three of this work; and Dr. W. Scott Watkins and Dr. Lynn B. Jorde for their contribution to chapter two and three of this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Long INterspersed Elements (LINE or L1) have had an enormous influence on human genomic structure, comprising about 20% of the mass of the human genome. In this analysis, the most recent L1 insertions in the human genome belonging to L1Hs Ta and preTa subfamilies were examined to further understand the impact L1 elements have had on human genomic structure and diversity. Collectively, over 800 human specific L1 elements from the draft sequence of the human genome were characterized. Estimates suggest that human specific L1 mobilization alone is responsible for increasing the size of the human genome by roughly 1.4 million bases, and that over 70 human specific L1 elements may still possess the ability to retrotranspose within human cells. Interestingly, over 35 L1 insertions were found adjacent to exons, though the majority of insertions showed general preference for gene poor regions of the genomes with low GC content. Analysis of over 500 L1 insertions by PCR on a diverse panel of humans representing geographically distinct human populations revealed that 115 (45%) of the Ta and 33 (14%) of the preTa human specific L1 insertions were variable in the human population with respect to insertion presence or absence. Sequence analysis of L1Hs Ta and preTa subfamily members yielded estimated average ages of 1.99 and 2.34 million years respectively. The 148 newly identified L1 insertion polymorphisms will serve as useful genetic markers for the study of human population genetics.

CHAPTER ONE:

BACKGROUND

Mobile elements were first identified as the cause of commonly observed chromosome break points at the Dissociation locus in the maize genome (McClintock 1956; McClintock 1987) and are most simply defined as linear DNA segments that can or could at one time move within a genome. Although some mobile elements are similar to viruses (Xiong and Eickbush 1988), they do not normally move from cell to cell. Mobile elements have had interesting and diverse histories within genomes. Although the origin and relationship between mobile elements and the genome in which they reside is a matter of debate, most often the function of mobile elements is unknown and probably not conserved from one species nuclear genome to another (Brosius 1991; Makalowski 2000). However, in rare cases they appear to serve a function, such as the Het-A and Tart elements which have been co-opted to maintain the integrity of telomeres in the *Drosophila* genome (Levis et al. 1993; Sheen and Levis 1994). Despite their lack of conserved function, mobile elements are present in almost every genome studied to date and are firmly established residents in the tree of life.

All genomes are organized differently with respect to their mobile element content. The pufferfish (*Fugu rubripes*) genome contains very few (<3%) mobile element derived repeats (Aparicio et al. 2002), whereas the fruit fly (*Drosophila melanogaster*) and the worm (*Caenorhabditis elegans*) genomes contain ~3% and ~10% mobile element derived repeats, respectively (The *C.elegans* Genome Consortium 1998; Myers et al. 2000; International Human Genome Sequencing Consortium 2001). By contrast, mammalian genomes are much more repeat rich, with over 37% of the mouse (*Mus musculus*) genome and 45% of the human genome composed of mobile elements (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002).

In addition to the differences in mobile element content, each genome has different types of mobile elements that can be can be broadly grouped into two major classes, DNA

mobile elements and retrotransposable elements, based on their mechanism of mobilization. In general, DNA transposons move in nonreplicative manner by a "cut and paste" mechanism in which the element is excised from one location and is inserted into another location, with the assistance of transposase protein encoded by the element (Mizuuchi 1992; van Luenen et al. 1994). Retrotransposons move by a "copy and paste" mechanism in which an RNA intermediate is generated (Weiner et al. 1986). This intermediate is then reverse transcribed by a reverse transcriptase into a cDNA that integrates into a new genomic location (Luan et al. 1993). The major type of mobile elements found in mammalian genomes is the retrotransposons (Deininger 1993b).

The most abundant type of mobile element in the human genome is a Short INterspersed Element (SINE) known as Alu (International Human Genome Sequencing Consortium 2001). Full length Alu elements are 300 bp in length, are dimeric in structure with a right and a left half, and are derived from the 7SL gene, a part of the signal recognition molecule (Rubin et al. 1980; Ullu and Tschudi 1984). Alu elements are RNA polymerase III transcribed retrotransposons and contain genomic encoded A rich tails (Rogers 1983). Alu elements are non-autonomous mobile elements, meaning they do not encode any proteins and are not capable of directing their own mobilization. Most Alu amplification occurred roughly 40-60 million years ago, with Alu elements growing in copy number to over a million in the haploid human genome and now representing ~10% of its total mass (Deininger and Daniels 1986; International Human Genome Sequencing Consortium 2001; Batzer and Deininger 2002).

Long INterspersed Elements (LINEs) are the other dominant mobile element family in the human genome (Singer 1982; Singer et al. 1993). LINEs are autonomous retrotransposons, meaning they encode all of the proteins required for their own mobilization.

Furthermore, it is believed that in order to move, Alu elements must scavenge the necessary proteins from LINE elements, therefore defining LINEs as the drivers of mobile element evolution in mammalian genomes (Jurka 1997; Kajikawa and Okada 2002). Full-length LINE elements are roughly 6000 base pairs (bp) in length and contain 5′ and 3′ untranslated regions (UTR) with an internal RNA polymerase II promoter (Prak and Kazazian 2000). They also contain two open reading frames (ORFs), both of which are required for retrotransposition, that are separated by a highly conserved 66 bp intergenic spacer (Fanning and Singer 1987; Singer et al. 1993; Moran et al. 1996). The first ORF encodes an RNA/DNA binding protein of unknown function (Martin and Bushman 2001). The second ORF encodes a reverse transcriptase with endonuclease activity (Sakaki et al. 1986; Mathias et al. 1991; Luan et al. 1993; Feng et al. 1996). In addition, LINEs usually have variable A rich tails at their 3′ ends. The youngest LINE class, LINE-1 or L1 elements, are believed to have arisen around the time of the mammalian radiation ~120 million years ago (Pascale et al. 1990; Smit 1996). They exist at a copy number of over 500,000 in the haploid human genome and account for 21% of its total mass (International Human Genome Sequencing Consortium 2001).

L1 elements mobilize using a process termed target primed reverse transcription (TPRT) (Luan et al. 1993). The process is initiated by an endonuclease generated nick at a target site that frees a short oligonucleotide sequence with a 3′ hydroxyl. This short oligonucleotide sequence with a 3′ hydroxyl serves to prime reverse transcription at the target site, with the L1 RNA serving as the template. The initial cleavage and reverse transcription event is followed by a second strand break of the DNA at the target site downstream of the initial endonuclease cleavage event. The process is completed by a second-strand DNA synthesis, ligation, and filling in of genomic sequence flanking the L1. This mechanism of

integration generates target site duplication sequences that can be used as useful landmarks defining the boundaries of newly integrated elements (Fanning and Singer 1987).

L1 elements have had a dramatic influence on the architecture of the human genome through a variety of mechanisms. First, L1 elements have greatly expanded the size of the human genome both, by their own retrotransposition and by providing the machinery necessary for the retrotransposition of other mobile elements, such as Alu elements (Jurka 1997; Kajikawa and Okada 2002). Second, L1 elements can disrupt normal gene function and have resulted in various human disease phenotypes, including Factor VIII hemophilia and Duchenne muscular dystrophy, by inserting within splice junctions, regulatory regions, or introns (Kazazian et al. 1988; Narita et al. 1993). Third, L1 elements can copy and carry non-L1 derived sequences throughout the genome in a unique L1-mediated duplication event known as three prime transduction (Moran et al. 1999; Goodier et al. 2000). Fourth, like other mobile elements, L1 elements can facilitate evolutionary rearrangements by providing sequence homology blanketing the genome which can be exploited during recombination (Nicholls et al. 1987; Schwartz et al. 1998; Hughes and Coffin 2001). More recently, L1 elements have also been shown to have paradoxical roles in genomic stability, by serving both as molecular band-aids in the repair of double stranded DNA breaks and as suspects for the generation of genomic deletions (Gilbert et al. 2002; Kazazian and Goodier 2002; Morrish et al. 2002).

Like other mobile elements, L1 and Alu elements also generate genetic differences both within and between primate species that can be analyzed using a simple polymerase chain reaction (PCR) assay, making them useful as genetic markers (Furano and Usdin 1995; Shimamura et al. 1997; Shedlock and Okada 2000). In addition, L1 and Alu elements are stable insertions that are rarely removed, the ancestral state is known to be the absence of the

element, they have only two alleles with regards to the presence or absence of the element, and they are variable in the human (Batzer and Deininger 1991; Perna et al. 1992; Batzer et al. 1994; Arcot et al. 1996), all of which makes mobile element insertion polymorphisms useful as unique genetic markers (Jorde et al. 2000; Roy-Engel et al. 2001; Watkins et al. 2001; Batzer and Deininger 2002). Unlike other genetic systems, most mobile element insertions are identical by descent, meaning that if two individuals share a mobile element insertion, those individuals most likely inherited the insertion from a common ancestor (Batzer et al. 1994; Hillis 1999). The essentially homoplasy-free nature of mobile element insertions makes them a rich source of population specific genetic variation (Roy-Engel et al. 2001; Batzer and Deininger 2002).

Despite the abundance of L1 elements in the human genome, only a limited number of L1 elements have been capable of mobilization throughout primate evolution under the "master gene" or limited amplification model (Deininger et al. 1992). In this model, L1 elements amplified from a series of "source genes" that gradually accumulated new mutations or "diagnostic" base substitutions. This gave rise to a series of L1 sequences with common nucleotide differences that comprise subfamilies or clades. In addition, the time frame over which a source gene was active has generated subfamilies with different genetic ages, based upon random, or non-diagnostic mutations that occur post integration (Deininger et al. 1992; Batzer et al. 1993; Deininger 1993a; Smit et al. 1995). In the human genome, the most recently integrated human L1 elements were identified as a result of their sequence identity to known retrotransposition competent elements and disease-causing *de novo* insertions, (Kazazian et al. 1988; Woods-Samuels et al. 1989; Miki et al. 1992; Schwahn et al. 1998; Meischl et al. 2000) and belong to the L1 Human specific Ta (transcribed subset a) and L1 preTa subfamilies (Skowronski et al. 1988; Boissinot et al. 2000)

Here, we determine the impact that L1 retrotransposition has had on human genomic structure and the amount of human genetic diversity associated with L1Hs Ta and preTa elements. Using the established sequence structure of L1 elements, specific oligonucleotide sequences containing diagnostic mutations unique to the Ta and preTa subfamilies were designed (Boissinot et al. 2000; Myers et al. 2002; Salem et al. 2003). These oligonucleotides were used to search the draft human genomic sequence to identify all of the Ta and preTa L1 elements in the available sequence of the human genome. Newly identified Ta and preTa elements were analyzed for sequence content and associated genetic variation, using L1 element specific PCR assays on diverse human and non-human primate DNA samples (Sheen et al. 2000; Myers et al. 2002; Salem et al. 2003). Using this approach, 830 recently integrated L1 elements from the human genome have been characterized. The newly identified L1 elements will provide useful genomic landmarks for primate comparative genomics and human population genetics.

References

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297:1301-1310

Arcot SS, Adamson AW, Lamerdin JE, Kanagy B, Deininger PL, Carrano AV, Batzer MA (1996) Alu fossil relics--distribution and insertion polymorphism. Genome Res 6:1084-1092

Batzer MA, Deininger PL (1991) A human-specific subfamily of Alu sequences. Genomics 9:481-487

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nat Rev Genet 3:370-379

Batzer MA, Schmid CW, Deininger PL (1993) Evolutionary analyses of repetitive DNA sequences. Methods Enzymol 224:213-232

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ, et al. (1994) African origin of human-specific polymorphic Alu insertions. Proc Natl Acad Sci U S A 91:12288-12292

Boissinot S, Chevret P, Furano AV (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17:915-928

Brosius J (1991) Retroposons--seeds of evolution. Science 251:753

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. Science 282:2012-2018

Deininger PL (1993a) Induction of DNA rearrangement and transposition. Proc Natl Acad Sci U S A 90:3780-3781

Deininger PL (1993b) Evolution of Retrotransposons. In: Evolutionary Biology. Hect M, MacIntyre RJ, Clegg M (eds). Vol 27. Plenum Publishing Corporation, New York, pp 157-196

Deininger PL, Batzer MA, Hutchison CA, 3rd, Edgell MH (1992) Master genes in mammalian repetitive DNA amplification. Trends Genet 8:307-311

Deininger PL, Daniels GR (1986) The recent evolution of mammalian repetitive DNA elements. Trends Genet 2:76-80

Fanning TG, Singer MF (1987) LINE-1: a mammalian transposable element. Biochim Biophys Acta 910:203-212

Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905-916

Furano AV, Usdin K (1995) DNA "fossils" and phylogenetic analysis. Using L1 (LINE-1, long interspersed repeated) DNA to determine the evolutionary history of mammals. J Biol Chem 270:25301-25304

Gilbert N, Lutz-Prigge S, Moran J (2002) Genomic Deletions Created upon LINE-1 Retrotransposition. Cell 110:315-325

Goodier JL, Ostertag EM, Kazazian HH, Jr. (2000) Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum Mol Genet 9:653-657

Hillis DM (1999) SINEs of the perfect character. Proc Natl Acad Sci U S A 96:9979-9981

Hughes JF, Coffin JM (2001) Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet 29:487-489

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979-988

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 94:1872-1877

Kajikawa M, Okada N (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell 111:433-444

Kazazian H, Goodier J (2002) LINE Drive. Retrotransposition and Genome Instability. Cell 110:277-280

Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332:164-166

Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM (1993) Transposons in place of telomeric repeats at a Drosophila telomere. Cell 75:1083-1093

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72:595-605

Makalowski W (2000) Genomic scrap yard: how genomes utilize all that junk. Gene 259:61-67

Martin SL, Bushman FD (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol 21:467-475

Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD, Gabriel A (1991) Reverse transcriptase encoded by a human transposable element. Science 254:1808-1810

McClintock B (1956) Controlling elements and the gene. Cold Spring Harbor Symp Quant Biol 21:197-216

McClintock B (1987) The discovery and characterization of mobile elements. The collected papers of Barbara McClintock. Garland, New York

Meischl C, Boer M, Ahlin A, Roos D (2000) A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. Eur J Hum Genet 8:697-703

Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res 52:643-645

Mizuuchi K (1992) Transpositional recombination: mechanistic insights from studies of mu and other elements. Annu Rev Biochem 61:1011-1051

Moran JV, DeBerardinis RJ, Kazazian HH, Jr. (1999) Exon shuffling by L1 retrotransposition. Science 283:1530-1534

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr. (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87:917-927

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31:159-165

Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, et al. (2000) A whole-genome assembly of Drosophila. Science 287:2196-2204

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA (2002) A Comprehensive Analysis of Recently Integrated Human Ta L1 Elements. Am J Hum Genet 71:312-326

Narita N, Nishio H, Kitoh Y, Ishikawa Y, Minami R, Nakamura H, Matsuo M (1993) Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. J Clin Invest 91:1862-1867

Nicholls RD, Fischel-Ghodsian N, Higgs DR (1987) Recombination at the human alpha-globin gene cluster: sequence features and topological constraints. Cell 49:369-378

Pascale E, Valle E, Furano AV (1990) Amplification of an ancestral mammalian L1 family of long interspersed repeated DNA occurred just before the murine radiation. Proc Natl Acad Sci U S A 87:9481-9485

Perna NT, Batzer MA, Deininger PL, Stoneking M (1992) Alu insertion polymorphism: a new type of marker for human population studies. Hum Biol 64:641-648

Prak ET, Kazazian HH, Jr. (2000) Mobile elements and the human genome. Nat Rev Genet 1:134-144

Rogers J (1983) Retroposons defined. Nature 301:460

Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, Batzer MA, Deininger PL (2001) Alu insertion polymorphisms for the study of human genomic diversity. Genetics 159:279-290

Rubin CM, Houck CM, Deininger PL, Friedmann T, Schmid CW (1980) Partial nucleotide sequence of the 300-nucleotide interspersed repeated human DNA sequences. Nature 284:372-374

Sakaki Y, Hattori M, Fujita A, Yoshioka K, Kuhara S, Takenaka O (1986) The LINE-1 family of primates may encode a reverse transcriptase-like protein. Cold Spring Harb Symp Quant Biol 51 Pt 1:465-469

Salem AH, Myers JS, Otieno AC, Scott Watkins W, Jorde LB, Batzer MA (2003) LINE-1 preTa Elements in the Human Genome. J Mol Biol 326:1127-1146

Schwahn U, Lenzner S, Dong J, Feil S, Hinzmann B, van Duijnhoven G, Kirschner R, Hemberger M, Bergen AA, Rosenberg T, Pinckers AJ, Fundele R, Rosenthal A, Cremers FP, Ropers HH, Berger W (1998) Positional cloning of the gene for X-linked retinitis pigmentosa 2. Nat Genet 19:327-332

Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D, Disteche C, McGillivray B, de la Chapelle A, Page DC (1998) Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Hum Mol Genet 7:1-11

Shedlock AM, Okada N (2000) SINE insertions: powerful tools for molecular systematics. Bioessays 22:148-160

Sheen FM, Levis RW (1994) Transposition of the LINE-like retrotransposon TART to Drosophila chromosome termini. Proc Natl Acad Sci U S A 91:12510-12514

Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD (2000) Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res 10:1496-1508

Shimamura M, Yasue H, Ohshima K, Abe H, Kato H, Kishiro T, Goto M, Munechika I, Okada N (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature 388:666-670

Singer MF (1982) SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28:433-434

Singer MF, Krek V, McMillan JP, Swergold GD, Thayer RE (1993) LINE-1: a human transposable element. Gene 135:183-188

Skowronski J, Fanning TG, Singer MF (1988) Unit-length line-1 transcripts in human teratocarcinoma cells. Mol Cell Biol 8:1385-1397

Smit AF (1996) The origin of interspersed repeats in the human genome. Curr Opin Genet Dev 6:743-748

Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246:401-417

Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. Nature 312:171-172

van Luenen HG, Colloms SD, Plasterk RH (1994) The mechanism of transposition of Tc3 in C. elegans. Cell 79:293-301

Watkins WS, Ricker CE, Bamshad MJ, Carroll ML, Nguyen SV, Batzer MA, Harpending HC, Rogers AR, Jorde LB (2001) Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. Am J Hum Genet 68:738-752

Weiner AM, Deininger PL, Efstratiadis A (1986) The reverse flow of genetic information : Pseudogenes and transposable elements derived from nonvira cellular RNA. Annu Rev Biochem 55:631-661

Woods-Samuels P, Wong C, Mathias SL, Scott AF, Kazazian HH, Jr., Antonarakis SE (1989) Characterization of a nondeleterious L1 insertion in an intron of the human factor VIII gene and further evidence of open reading frames in functional L1 elements. Genomics 4:290-296

Xiong Y, Eickbush TH (1988) Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol 5:675-690

CHAPTER 2:

A COMPREHENSIVE ANALYSIS OF RECENTLY INTEGRATED HUMAN Ta L1 ELEMENTS*

*Reprinted by permission of The American Journal of Human Genetics

Introduction

Computational analysis of the draft sequence of the human genome indicates that repetitive sequences comprise 45-50% of the human genome mass, 17% of which consists of ~500,000 L1 LINEs (Long INterspersed Elements) (Smit 1999; Prak and Kazazian 2000; Lander et al. 2001). L1 elements are restricted to mammals, having expanded as a repeated DNA sequence family over the past 100-150 million years (Smit et al. 1995). Full-length L1 elements are ~6kb long and amplify via an RNA intermediate in a process known as retrotransposition. L1 integration likely occurs by a mechanism termed target primed reverse transcription (TPRT) (Luan et al. 1993; Kazazian and Moran 1998). This mechanism of mobilization provides two useful landmarks for the identification of L1Hs (L1 human specific) inserts: an endonuclease related cleavage site (Jurka 1997; Cost and Boeke 1998; Cost et al. 2001) and direct repeats or target site duplications flanking newly integrated elements (Fanning and Singer 1987; Kazazian 2000).

L1 retrotransposons have had a significant impact on the human genome, through recombination (Fitch et al. 1991), alteration of gene expression (Yang et al. 1998; Rothbarth et al. 2001), and *de novo* insertions that disrupt open reading frames (ORFs) and splice sites resulting in human disease (Kazazian et al. 1988; Kazazian 1998; Kazazian and Moran 1998). L1 elements are also able to transduce adjacent genomic sequences at their 3′ end, facilitating exon shuffling (Boeke and Pickeral 1999; Moran et al. 1999; Goodier et al. 2000). In addition, individual mobile elements may undergo post-integration gene conversion events in which short DNA sequences are exchanged by an undefined mechanism, thereby altering the levels of single nucleotide polymorphism (SNP) associated with the individual L1 elements (Hardies et al. 1986). Thus, LINEs have exerted a significant influence on the architecture of the human genome.

Even though there are ~500,000 L1 elements in the human genome, only a limited subset of L1 elements appear to be capable of retrotransposition (Moran et al. 1996; Sassaman et al. 1997). As a result of the limited amplification potential of this diverse gene family, a series of discrete subfamilies of L1 elements exists within the human genome (Deininger et al. 1992; Smit et al. 1995). Each of the L1 subfamilies appears to have amplified at different times in primate evolution, making them different ages (Deininger et al. 1992; Smit et al. 1995). The most recently integrated L1 elements within the human genome share a common 3-bp diagnostic sequence within the 3′ untranslated region (UTR), and they comprise almost all of the *de novo* disease-associated L1 elements within the human genome, as well as several elements that have been shown to be capable of retrotransposition in cell culture (Kazazian and Moran 1998; Boissinot et al. 2000; Sheen et al. 2000). This subfamily was first identified in human teratocarcinoma cells and has been collectively termed "Ta" (for transcribed, subset a) (Skowronski et al. 1988). Some members of the L1 Ta subfamily have inserted in the human genome so recently that they are polymorphic with respect to insertion presence/absence (Boissinot et al. 2000; Sheen et al. 2000). The L1 insertion polymorphisms are a useful source of identical-by-descent variation for the study of human population genetics (Boissinot et al. 2000; Santos et al. 2000; Sheen et al. 2000). Here, we report the analysis of the Ta subfamily of L1 elements from the draft sequence of the human genome.

Materials and Methods

Cell Lines and DNA Samples

The cell lines used to isolate primate DNA samples were as follows: human (*Homo sapiens*) HeLa (American Type Culture Collection [ATCC] number CCL2), common chimpanzee (*Pan troglodytes*) Wes (ATCC number CRL1609), pygmy chimpanzee (*Pan paniscus*) (Coriell Cell Repository number AG05253), gorilla (*Gorilla gorilla*) Lowland Gorilla

(Coriell Cell Repository number AG05251B), green monkey (*Cercopithecus aethiops*) (ATCC number CCL70), and owl monkey (*Aotus trivirgatus*) (ATCC number CRL 1556). Cell lines were maintained as directed by the source, and DNA isolations were performed using Wizard genomic DNA purification (Promega). Human DNA samples from the European, African American, Asian or Alaskan native, and Egyptian population groups were isolated from peripheral blood lymphocytes (Ausabel et al. 1987), as described elsewhere (Stoneking et al. 1997).

Computational Analyses

The draft sequence of the human genome was screened using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990), available at the National Center for Biotechnology Information genomic BLAST Web site. A 19-bp oligonucleotide (5′-CCTAATGCTAGATGACACA-3′) that is diagnostic for the L1Hs Ta subfamily was used to query the human genome database with the following the optional parameters: filter none and advanced options -e 0.01, -v 600, -b 600. Copy number estimates were determined from BLAST search results. Sequences containing exact matches were subjected to additional analysis as outlined below.

A sequence region of 9,000-10,000 bp, including the match and 1,000-2,000 bp of flanking unique sequence, was annotated using RepeatMasker (version 7/16/00), from the University of Washington Genome Center, or Censor, from the Genetic Information Research Institute (Jurka et al. 1996). These programs annotate repeat sequence content and were used to confirm the presence of L1Hs elements and regions of unique sequence flanking the elements. Polymerase-chain-reaction (PCR) primers flanking each L1 element were designed using Primer3 software, available at the Whitehead Institute for Biomedical Research, and were complementary to the unique sequence regions flanking each L1 element. The resultant primers

were screened, by standard nucleotide-nucleotide BLAST (blastn) against the non-redundant (nr) and high-throughput (htgs) sequence databases, to ensure they resided in unique DNA sequences. Primers that resided in repetitive sequence regions were discarded, and, if possible, new primers were then designed. A complete list of all the L1 elements that were identified using this approach and supplemental material from this manuscript are available from the Batzer Lab Web site, in the "Publications" section. This supplementary data are also found as an appendix to this work. Individual L1 DNA sequences were aligned using MegAlign, with the Clustal V algorithm and the default settings (DNAstar, version 5.0 for Windows), followed by manual refinement.

PCR Amplification

PCR amplification of 262 individual L1 elements was performed in 25-µl reactions that contained 50-100 ng of template DNA; 40 pmol of each oligonucleotide primer (see Table 2.1; APPENDIX B Supplementary Data Table 1); 200 µM of deoxyribonucleoside triphosphates, in 50 mM KCl and 10 mM Tris-HCl (pH 8.4); 1.5 mM $MgCl_2$; and 1.25 U of Taq DNA polymerase. Each sample was subjected to the following amplification conditions for 32 cycles: an initial denaturation at 94 °C for 150 s, 1 min denaturation at 94 °C, and 1 min at the annealing temperature (specific for each locus, as shown in Table 2.1 and APPENDIX B supplementary Data Table 1), followed by extension at 72 °C for 10 min. For analysis, 20 µl of each sample was fractionated on a 2% agarose gel with 0.05 µg/ml ethidium bromide. PCR products were directly visualized using UV fluorescence. The human genomic diversity associated with each Ta L1 element was determined by the amplification of 20 individuals from each of four geographically distinct populations (African-American, Asian or Alaskan Native, European German, and Egyptian).

Table 2.1 - Summary of Ta L1 Computational and PCR Analysis[1]

Classification	Number of Elements
Successful PCR analysis	262
L1 elements inserted in other repeats	137
L1 elements located at the end of sequencing contigs	69
Total Ta L1 elements analyzed	468

[1] A full summary of GenBank accession numbers, PCR primers and conditions, and PCR amplicon sizes for these loci are shown in the APPENDIX B Supplementary Data Table 1.

Cloning and Sequence Analysis

L1 element-related PCR products were cloned using the Invitrogen TOPO TA Cloning® Kit, according to the manufacturer's instructions, and were sequenced using an Applied Biosystems 3100 automated DNA sequencer, by the chain-termination method (Sanger et al. 1977). The DNA sequence for the common and pygmy chimpanzee orthologs of L1HS72 were assigned GenBank accession numbers AF489459 and AF489460, respectively. Additional diverse human sequences from L1HS72 were assigned GenBank accession numbers AF489450-AF489458. DNA sequences derived from L1 pre-integration sites were assigned accession numbers AF461364, AF461365, AF461368-AF461383, AF461386, and AF461387.

Results

L1 Ta Subfamily Copy Number and Age

To identify recently integrated Ta L1 elements in the human genome, we searched the draft sequence of the human genome (BLASTN database, version 2.2.1), using BLAST (Altschul et al. 1990) with an oligonucleotide that is complementary to a highly conserved sequence in the 3′ UTR of Ta L1 elements. This 19-bp query sequence (CCTAATGCTAGATGACACA) includes the Ta subfamily-specific diagnostic mutation ACA at its 3′ end at positions 5930-5932 relative to L1 retrotransposable element-1 (Dombroski et al. 1991). We identified 468 unique Ta

L1 elements from 2.868 x10^9 bp of available human draft sequence. Extrapolating this number to the actual size of the human genome (3.162 x10^9 bp) and assuming the available sequence data is representative of the entire human genome, we estimate that this subfamily contains ~520 elements. Of the 468 elements retrieved, 69 resided at the end of sequence contigs and were not amenable to additional in vitro wet-bench analysis. Of the 399 remaining elements, 124 (31%) of the elements were essentially full length, and the remaining 275 were truncated to variable lengths. Alignment and sequence analysis of the full-length elements revealed that 44 contained two intact ORFs and therefore may be capable of retrotransposition. This estimate of putative retrotransposition-competent L1 elements is in good agreement with the initial analysis of the draft sequence of the human genome (Lander et al. 2001).

The ages of L1 elements can be determined by the level of sequence divergence from the subfamily consensus sequence by use of a neutral mutation rate for primate noncoding sequence of 0.15% per million years (Miyamoto et al. 1987). The mutation rate is known to be ~10 times greater for CpG bases as compared to non-CpG bases, as result of the spontaneous deamination of 5-methyl cytosine (Bird 1980). Thus, two age estimates that are based on CpG and non-CpG mutations can be calculated for the Ta subfamily of L1 elements. A total of 89,929 bp from the 3′ UTR of 459 Ta L1Hs elements were analyzed, and L1 elements characterized elsewhere were excluded from this analysis, as were nine elements that technically do not belong to the Ta subfamily according to the nucleotide present at position 6015 in the 3′ UTR of the elements (Ovchinnikov et al. 2001). Three hundred thirty-one total nucleotide substitutions were observed. Of these, 263 were classified as non-CpG mutations against the backdrop of 88,141 total non-CpG bases, thereby producing a non-CpG mutation density of 0.002984. Based on the non-CpG mutation density and a neutral rate of evolution (0.002984/0.0015), the average age of the Ta L1 elements was 1.99 million years. A total of 68 CpG mutations were found across these

459 L1 elements from 1,788 total CpG nucleotides, thereby yielding a CpG-mutation rate of 0.038031. With the expectation that the CpG mutation rate is ~10-fold higher than the non-CpG mutation rate, the approximate age (obtained using the CpG mutation density) of the L1Hs Ta subfamily is 2.54 million years. These estimates are in good agreement with one another, as well as with previous estimates derived from an analysis of a small number of Ta L1 elements (Boissinot et al. 2000).

Nine of the 468 elements analyzed do not technically belong to the Ta subfamily of L1 elements, on the basis of a single-nucleotide substitution (L1HS19, -72, -274, -309, -318, -325, -390, -399, and -493) that is also considered diagnostic for the L1 Ta subfamily. Although they all have the 19-bp query sequence ending in ACA in the 3′ UTR at positions 5930-5932, they lack a G at position 6015 (Ovchinnikov et al. 2001) and instead contain an A at that position, which is a diagnostic feature found in older primate-specific L1PA10-L1PA2 subfamilies (Smit et al. 1995). Thus, these elements may be Ta L1 elements that have undergone fortuitous single-base substitutions of the ancestral nucleotide, may be Ta L1 elements that have undergone backward gene conversion events, or may simply be older, "pre- Ta" L1 elements that were generated by a source gene (or source genes) that did not contain this diagnostic base. To determine the effect that the Ta versus non-Ta designation has on the calculated age estimate, we examined a total of 1,807 bp from the 3′ UTRs of these nine elements. There were 27 non-CpG mutations from a total of 1,771 non-CpG bases, thereby yielding a mutation density of 27/1,771, or 0.015246. Dividing by the neutral rate of evolution for primate noncoding sequence (0.015246/0.0015), we arrive at an estimated age of 10.16 million years. This is significantly older than the average age of 2.26 million years that was calculated from the larger data set (i.e., the data set of Ta L1 elements only including CpG and non-CpG mutations). The CpG mutation density in the elements was also calculated. There were 2 CpG mutations from 36 CpG bases,

thereby producing a CpG mutation density of 2/36, or 0.056. This figure was divided by the projected CpG mutation rate (0.056/.015), arriving at an estimated age of 3.73 million years. This figure is lower than the non-CpG mutation rate, but it still suggests that these elements are at least twice as old as their true Ta counterparts. In addition, all but one of these Ta L1 elements (L1HS493) was monomorphic for the presence of the L1 element in the human population. Thus, the higher levels of nucleotide diversity and absence of associated insertion polymorphism of eight of these L1 elements are consistent with their being older members of the L1 Ta subfamily, whereas L1HS493 may be the product of a gene conversion event.

The nucleotide-sequence substitution patterns were further examined with respect to the levels of presence/absence insertion polymorphism associated with each of the L1 elements (as outlined in detail below, in the "L1 Element-Associated Human Genomic Diversity" subsection). The 3′ UTRs of 139 fixed-present elements were analyzed for both CpG and non-CpG mutations and had an estimated average age of 2.45 million years. This calculation yields an age that is somewhat older than the average age predicted for the subfamily as a whole, a finding that was expected, since these elements are thought to have inserted during the early stages of L1Hs Ta expansion in the human genome, such that they have become fixed across diverse human populations. Similar calculations were repeated for the high-frequency, intermediate-frequency, and low-frequency L1 Ta insertion polymorphisms, with average ages of 2.24, 2.06, and 1.69 million years, respectively. Although the age differences across different insertion frequencies are not significantly different (P values >0.05) when tested with a one-tailed t test, they do suggest a progressive decrease in the calculated age of each group, with corresponding decreases in insertion frequency. This is exactly what would be expected under a model in which newer elements arose more recently and have lower allele frequencies in the human population.

L1 Associated Human Genomic Diversity

Of the 468 L1HS Ta elements isolated *in silico*, 262 were further analyzed using a PCR-based assay and flanking unique sequence primers as described elsewhere (Sheen et al. 2000) (Table 2.1 and APPENDIX B Supplementary Data Table 1). The remaining elements were not suitable for further analysis, for various reasons. Some (137) of the L1 elements were inserted into other repetitive regions of the genome such that flanking unique sequence PCR primers could not be designed. Sixty-nine additional elements resided at the end of sequencing contigs in GenBank, so the lack of flanking unique sequence information made PCR-primer design in this region impossible. Three elements, L1HS17, L1HS47 and L1HS63, produced inconclusive PCR results because of the amplification of paralogous genomic sequences as described elsewhere (Batzer et al. 1991). Another five elements produced non-specific PCR results, and they were excluded from further analysis. Thirty-six of the Ta L1 elements mapped to chromosome X, and 10 mapped to chromosome Y (Table 2.1 and APPENDIX B Supplementary Data Table 1). All of the Ta L1 elements from chromosomes X and Y were tested using human DNA samples in which the gender had been determined using a PCR-based assay that was described elsewhere (Eng et al. 1994). The human genomic diversity associated with the autosomal and sex-linked Ta L1 elements is summarized in Table 2.2 and APPENDIX B Supplementary Data Table 1 and 2.

A high degree (45%) of insertion polymorphism was found in the 254 (i.e., 262 - 8) remaining elements that were subjected to the two-step PCR-based assay across 80 individuals from four geographically diverse human populations (Table 2.2 and APPENDIX B Supplementary Data Table 2). One hundred thirty-nine of the Ta L1 elements were fixed present, meaning that every individual tested was homozygous (i.e., +/+) for the presence of the

L1 repeat. These elements are likely to be slightly older than their polymorphic counterparts, having inserted into the human genome prior to the migration of humans from Africa.

Table 2.2 - Summary of Ta L1 Element-Associated Human Genomic Diversity[1,2]

Autosomal Ta L1 elements	**Number of Elements**
High frequency insertion polymorphism	36
Intermediate frequency insertion polymorphism	55
Low frequency insertion polymorphism	15
Very low frequency / fixed absent	3
Fixed present	129
X-Linked Ta L1 elements	
High frequency insertion polymorphism	1
Intermediate frequency insertion polymorphism	1
Low frequency insertion polymorphism	4
Very low frequency / fixed absent	0
Fixed present	8
Y-Linked Ta L1 elements	
Polymorphic	0
Fixed present	2

[1] The Ta L1 insertion polymorphisms are classified according to allele frequency as: high-frequency (HF) (present in more than 2/3 but not in all chromosomes tested), intermediate-frequency (IF) (present in more than 1/3 of chromosomes tested but in no more than 2/3 of the chromosomes), low-frequency (LF) (present in no more than 1/3 of the chromosomes tested), or very-low-frequency (VLF) (or "private") insertion polymorphisms.

[2] A full summary of the genotypes for each locus, L1 allele-frequency data, and heterozygosity values are shown in the supplementary data and are also available from the Batzer laboratory web site.

By contrast, 115 of the elements assayed by PCR were polymorphic, to some degree, in the populations that were surveyed. A survey of human genomic diversity associated with a severely truncated L1 element is shown in figure 2.1. A sample of the human genomic diversity associated with relatively long L1 insertion polymorphism is shown in figure 2.2. Thirty-seven of the Ta L1 elements were high-frequency insertion polymorphisms with an L1 allele frequency that was >0.67, so that most of the individuals were homozygous for the presence of the L1

element. Fifty-six of the polymorphic elements were intermediate frequency, with an L1 allele frequency >0.33 but <0.67 across the diverse human populations sampled. Nineteen of the 254 elements had insertion allele frequencies <0.33, and these were termed "low-frequency insertion polymorphisms." These elements include some of the youngest members of the subfamily, having inserted into the human genome so recently that the element only appears in the genomes of only a handful of individuals who were screened by our assay. Three Ta L1 elements, L1HS44, L1HS287, and L1HS373, appeared to be absent from the genomes of all the individuals tested, and one of these (L1HS373) is full length and has two functional ORFs, suggesting that it may be retrotransposition competent. Previous experiments with Alu elements have shown that

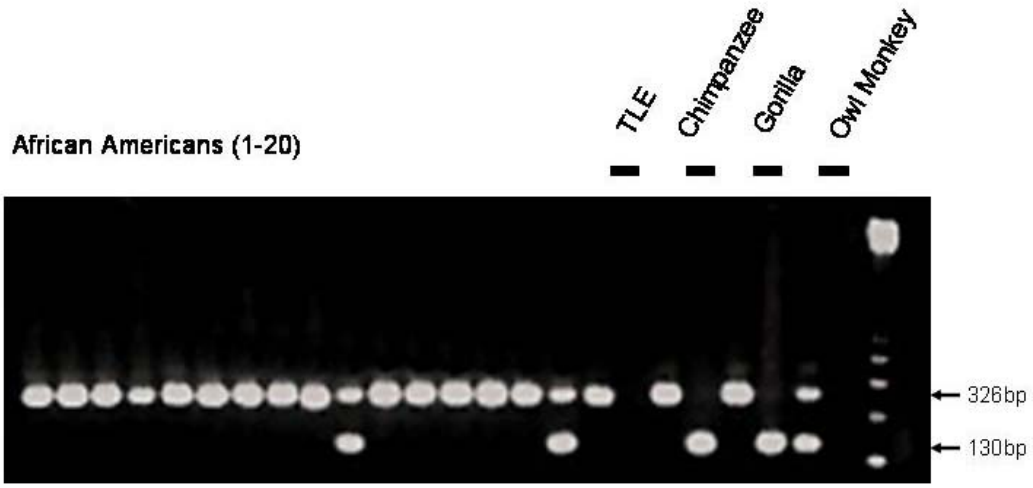

Figure 2.1

Figure 2.1 - Human diversity associated with a truncated Ta L1Hs element, as shown by an an agarose gel chromatograph of the PCR products from a survey of the human genomic variation associated with L1HS7. Amplification of the pre-integration site of this locus generates a 130-bp PCR product; amplification of a filled site generates a 326-bp product (by use of flanking unique sequence primers). In this survey of human genomic variation, 20 individuals from each of four diverse populations were assayed for the presence or absence of the L1 element, with only the African-American samples shown here; the control samples (*gray lines*) were TLE buffer (i.e., 10 mM Tris-HCl:0.1 mM EDTA), common chimpanzee, gorilla, and owl monkey DNA templates. Most of the individuals surveyed were homozygous for the presence of the L1 element; in addition, this particular L1 element was absent from the genomes of nonhuman primates.

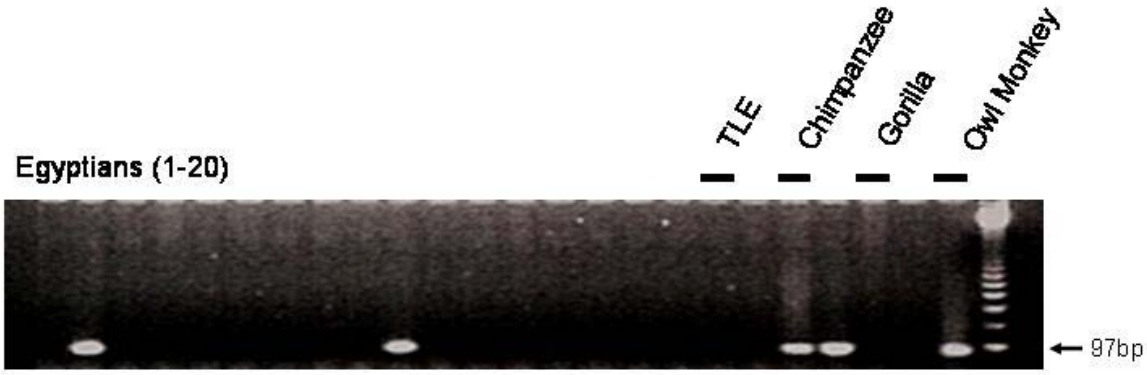

Figure 2.2A

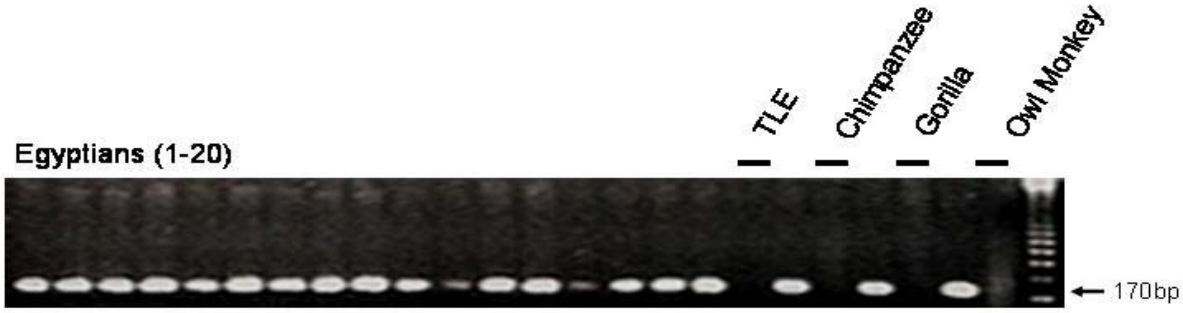

Figure 2.2B

Figure 2.2 - Human diversity associated with a long L1HS Ta insertion polymorphism, as shown by an agarose gel chromatograph of the PCR products from a survey of the human genomic variation associated with L1HS364. Because of the size (~6000 bp) of this L1 element, two separate PCRs are performed to genotype individual samples. In the first reaction, flanking unique sequence primers were used to genotype the empty alleles **(A)**; amplification of empty alleles from this locus generates a 97-bp PCR product. In the second reaction, a Ta subfamily-specific internal primer termed "ACA" and the 3′ flanking unique sequence primer were used to genotype filled sites **(B)**; the amplification of filled sites generates a 170-bp product. In this survey of human genomic variation, 20 individuals from each of four diverse populations were assayed for the presence or absence of the L1 element, with only the Egyptian samples shown here; the control samples (*black lines*) were TLE buffer, common chimpanzee, gorilla, and owl monkey DNA templates. This particular L1 insertion polymorphism is a high-frequency insertion polymorphism, and most of the individuals surveyed have L1 filled chromosomes.

these types of elements are indeed present within the genomic clone that was sequenced as part of the human genome project, but also that they represent relatively rare, "private" mobile-element insertion polymorphisms (Carroll et al. 2001).

Overall, the unbiased heterozygosity values across all of the L1 elements subjected to PCR analysis were similar across the four populations, with values of 0.265 in African Americans, 0.233 in Asians, 0.252 in European Germans (i.e., white Germans of European descent), and 0.250 in Egyptians (Table 2.2 and APPENDIX B Supplementary Data Table 2). However, several of the polymorphic elements individually exhibited unbiased heterozygosity values that approached 0.5, the theoretical maximum for bi-allelic loci. A subset of 31 of the 115 L1 insertion polymorphisms are, to some degree, population specific, meaning that insertion frequencies differ by at least 25% in one of the tested populations, relative to the other three populations that were surveyed. Detailed analysis of the human genomic variation associated with the polymorphic L1 elements will prove useful for the study of human population genetics.

In order to determine if the LINE insertion polymorphisms were in Hardy-Weinberg Equilibrium (HWE), we performed a total of 460 χ^2 tests for goodness of fit. A total of 77 deviations from Hardy-Weinberg expectations were observed in the comparisons. However, 73 of the deviations were the result of low expected numbers. The remaining four tests that deviated from HWE did not cluster by locus or population. A total of 23 deviations from HWE would be expected by chance alone at the 5% significance interval. In addition, we applied the Fisher's exact test to the data, using the Genetic Data Analysis program. The test yielded only 22 of 436 significant comparisons, which is approximately what would be expected on the basis of chance alone. By Fisher's exact test, only 6 of the 436 comparisons were significant at the 0.01 level, and they did not cluster across all populations at any locus tested. Therefore, we conclude that these L1 insertion polymorphisms do not significantly depart from HWE.

Phylogenetic Origin

Almost all of the Ta L1 elements analyzed using PCR were located in the human genome and absent from the orthologous positions within nonhuman primate genomes. Only a single truncated L1 element (L1HS72) produced unexpected results when subjected to the initial PCR by use of external flanking primers and nonhuman primate DNA as a template. The 825-bp amplicon that corresponded to the L1HS72 insertion was found in loci in all 80 human individuals tested, as well as in the orthologous loci from the common chimpanzee and pygmy chimpanzee genomes (figure 2.3A). However, the gorilla, green monkey, and owl monkey only amplified the small PCR product corresponding to the empty allele or pre-integration site (figure 2.3A). Subsequent PCRs by use of the internal subfamily-specific ACA primer and the 3′ flanking primer across the same DNA templates produced a characteristic L1 filled-site amplicon only in the human individuals and not in any of the nonhuman primate genomes (chimpanzee, gorilla, green monkey, and owl monkey). It appeared that we had potentially isolated a Ta L1 element that inserted into the genome before the divergence of humans from African apes, but the second PCR by use of the internal subfamily-specific ACA primer and the 3′ flanking primer again produced the expected product that corresponded to the presence of this Ta L1 element only in humans. These data suggest that there is a difference in the sequence structure of this L1 element in the human genome, as compared to the common and pygmy chimpanzee genomes, which contained putative Ta L1 filled alleles.

Gene Conversion

To precisely define the sequence structure of the L1HS72 locus, we cloned and sequenced, for further analysis, the PCR amplicons from several human genomes, as well as those from the common chimpanzee and the pygmy chimpanzee (figure 2.3B). Sequence analysis of the orthologous sites from the common and the pygmy chimpanzee genomes revealed

the presence of an older, primate-specific L1 element that had the greatest sequence identity to the L1PA3 subfamily (figure 2.3B). Interestingly, this L1 element shared identical target site duplications with that of the Ta L1 element that was present in the human samples that we studied. Both the human sequence and the chimpanzee sequence also contained many of the diagnostic mutations characteristic of an L1PA3 element. However, only the human L1 sequences contained the Ta diagnostic ACA mutation at position 5930-5932 in the 3′ UTR. The common and pygmy chimpanzee sequences contained GAT at this position and an additional A mutation at diagnostic position 6015, both of which are characteristic of older L1PA elements (L1PA6 –L1PA2). The most likely explanation for the presence of the L1Hs Ta ACA sequence in the human L1 element is a forward gene conversion event that affected a pre-existing older L1 element at this locus. To further investigate the putative gene conversion at this locus, we cloned and sequenced alleles derived from African American, Asian, European German, and Egyptian genomes. Although there was a limited sample size, all nine individuals who were sequenced contained the ACA sequence, and at least four samples (European Germans 1 and 2, and Egyptians 2 and 3) contained SNPs, three of which occur at a specific CpG dinucleotide (figure 3B). Therefore, we conclude that gene conversion events have altered the L1 Ta subfamily-specific diagnostic nucleotide positions at
this locus within the human lineage.

In order to begin to examine the level of gene conversion across the entire Ta subfamily, we examined multiple sequence alignments of the 459 Ta L1Hs elements. Close inspection of the multiple sequence alignment revealed some highly variable sequence features that were unexpected among such a young L1 subfamily, in which we would expect low levels of nucleotide divergence. It appears that many of the single-base substitutions in Ta L1 elements are not completely random mutation events. In fact, it became clear that a substantial number of

the elements possessed specific mutations that are diagnostic for older L1PA primate-specific elements in addition to the younger diagnostic mutations. These mosaic elements all possessed the 19-bp Ta L1 consensus sequence, but they also contained short tracts of sequence diagnostic for other L1 subfamilies.

There are two possible explanations for the presence of these mosaic elements. The first theory is that L1Hs Ta source genes, while acquiring the young diagnostic mutations of the L1Hs Ta subfamily, also retained many of the other diagnostic mutations of their older L1 subfamily progenitors. Over time, this gave rise to elements with combinations of young and old mutations, as proposed in the master-gene theory of LINE and short interspersed element (SINE) amplification (Deininger et al. 1992). The second theory is that some of these mosaic elements are products of gene conversion events, a non-reciprocal transfer of sequence between a pair of nonallelic genomic DNA sequences such as interspersed repeats. The donor sequence is unchanged, and the recipient sequence gains some of the donor sequence; alternatively, a nonintegrated LINE cDNA may also serve as the donor sequence for the gene conversion. Gene conversion between SINEs and LINEs is a significant influence on the genomic landscape of young Alu elements, creating hybrid sequence mosaics of the various mobile element subfamilies (Batzer et al. 1995; Kass et al. 1995; Roy et al. 2000; Roy-Engel et al. 2001; Roy-Engel et al. 2002). Gene conversion may contribute to as much as 10-20% of the sequence variation between recently integrated Alu elements (Roy et al. 2000). It is likely that the same process may also alter the sequence diversity of L1 elements, since they are also part of a large, nearly identical multigene family and since they have previously been shown to have undergone limited gene conversion (Hardies et al. 1986; Burton et al. 1991). Unfortunately, the vast majority of primate L1 subfamily structure has only been deduced computationally and has not

Figure 2.3 - L1HS72 gene conversion. A) Agarose gel chromatograph of the PCR products derived from the amplification of L1HS72 in a series of human and nonhuman primate genomes, with a schematic of the primate evolutionary tree over the past 35 million years shown below. The yellow notched arrow represents the approximate time period when the L1HS72 element first integrated, and the red notched arrow represents the approximate time period of the gene conversion event of the pre-existing L1 element. The fragment-length marker is a 123-bp ladder. **B)** Sequence alignment generated by sequencing the L1HS72 amplicons from nine diverse humans. Sequences are compared relative to L1Hs Ta consensus sequence and the L1HS72 sequence obtained from GenBank with only the diagnostic bases shown and positions reported relative to L1 retrotransposable element-1 (Dombroski et al. 1991). The G and C at positions 5536 and 5539 are indicative of the Ta-0 subset, whereas the Ta-1 subset has T and G at these nucleotides (Boissinot et al. 2000). The G at position 6015 (in addition to the ACA at positions 5930-5932) is diagnostic for the L1Hs Ta subfamily (Ovchinnikov et al. 2001). The target site duplication sequence (TSD) is shown in brackets. The mosaic elements seen in the human samples are believed to be the result of at least one gene conversion, some time after the divergence of humans from the great apes (approximately 5 million years ago), of a pre-existing L1 element with a younger L1Hs element. In the representation of nucleotides, different colors are used to denote conserved sequences and sequence variations between samples: green denotes bases unique to the common and pygmy chimpanzee genomes; blue denotes nucleotides unique to the human samples; orange denotes shared bases conserved between the common chimpanzee, pygmy chimpanzee, and human samples; red denotes SNPs, within L1HS72, in the human population.

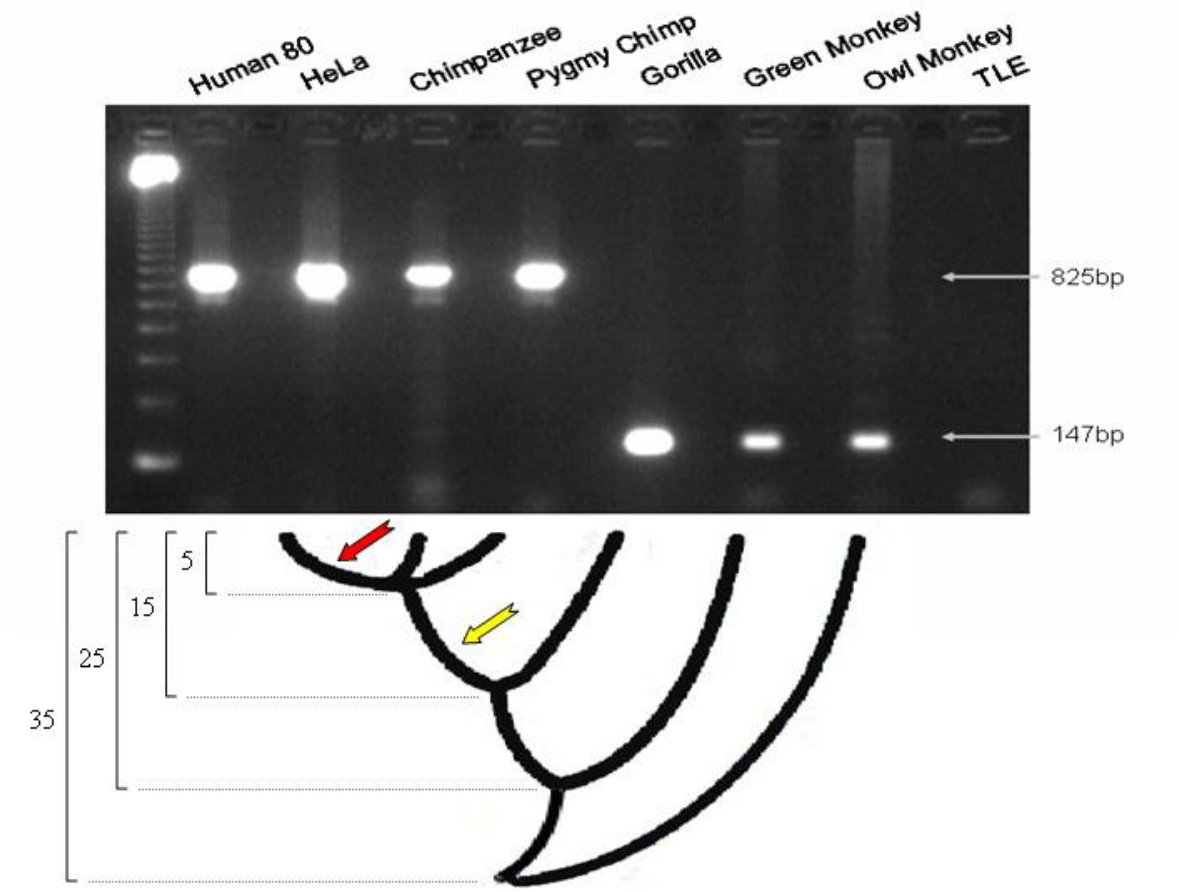

Figure 2.3A

Position information[1]	ORF 2 (5536,5539) (5591) (5713) (5745) (5767) (5791)	3′UTR (5828) (5851) (5864) (5871) (5879) (5930-5932) (5946) (5990) (6015)
	(AAGATTCTA)	(AAGATTCTA)
L1HS-Ta consensus	5′.....(T..G)*...C...A...C...G...C..|..T...T...G...G...C...ACA...CpG...A...G².....	3′
L1HS-Ta0 consensus	5′.....(G..C)*...C...A...C...G...C..|..T...T...G...G...C...ACA...CpG...A...G².....	3′
L1HS-Ta1 consensus	5′.....(T..G)*...C...A...C...G...C..|..T...T...G...G...C...ACA...CpG...A...G².....	3′
L1HS72 GenBank	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....TSD...	3′
African American 1	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....TSD...	3′
African American 2	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....TSD...	3′
Asian 1	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....TSD...	3′
Asian 2	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....TSD...	3′
German Caucasian 1	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....T.....TSD...	3′
German Caucasian 2	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....T.....TSD...	3′
Egyptian 1	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....TSD...	3′
Egyptian 2	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...G...T...G.....T.....TSD...	3′
Egyptian 3	5′...TSD...(G..C)*...T...G...T...A...T..|..A...C...A...T...G.....T.....T.....TSD...	3′
Chimpanzee	5′...TSD...(G..C)*...T...G...T.....T..|..A...C.....T...G...GAT.....TSD...	3′
Pygmy Chimpanzee	5′...TSD...(G..C)*...T...G...T.....T..|..A...C.....T...G...GAT.....TSD...	3′

Figure 2.3B

been verified at the wet bench, to precisely define the expansion of L1 elements in a phylogenetic context. Therefore, it is currently not possible to accurately estimate the level of gene conversion between L1 elements within the genome.

Sequence Diversity

One hallmark of L1 integration is the generation of target site duplications flanking newly integrated elements. Two thousand base pairs of flanking sequence on each side of the element were searched for target site duplications. Direct repeats >10 bp long are considered to be clear target site duplications. Of the 399 elements (i.e., a total of 468 elements minus the 69 elements located at the end of sequencing contigs), we were able to identify clear target site duplications for 272 elements. All elements with clear target site duplications had endonuclease sites that matched those described elsewhere (Feng et al. 1996; Jurka 1997; Cost and Boeke 1998). A total of 13 elements (L1HS45, -70, -172, -178, -284, -372, -415, -416, -442, -443, -448, -513, and -558) apparently lacked target site duplications or contained short target site duplications. To further investigate these elements, PCRs specific for the pre-integration sites for those elements listed were performed on the common chimpanzee, pygmy chimpanzee, and, when possible, human samples. The resulting amplicons were cloned and sequenced, to unambiguously define the pre-integration site for each element. The resulting pre-integration sites were then compared with the original GenBank sequence for each locus.

All 13 of the L1Hs elements lacked obvious target site duplications when compared with the common and pymgy chimpanzee pre-integration site sequences. In addition, L1HS178, and L1HS284 had no observable target site duplications and atypical endonuclease cleavage sites. One possible explanation for this observation is that these elements have integrated independent by endonuclease cleavage of target sequence; this mechanism has been proposed for the repair of doubled-stranded breaks in DNA (Moore and Haber 1996; Teng et al. 1996;

Morrish et al. 2002). Alternatively, these elements may represent forward gene conversion events of pre-existing L1 elements that, by mutation, have rendered their target site duplications unrecognizable. However, because little is known about the rates of these endonuclease-independent insertion events in mammalian cells, further studies are required in order to resolve the mechanism underlying these integration events.

Another aspect of L1Hs Ta sequence diversity is created by variable 5′ truncation such that some of the elements in the human genome are only a few hundred base pairs long, whereas some full-length elements are >6000 bp long. This phenomenon is classically attributed to the lack of processivity of the reverse transcriptase enzyme in the creation of the L1 cDNA copy. The point of truncation is traditionally believed to occur as a function of length, where shorter inserts are more likely to occur in the human genome than longer elements (Grimaldi et al. 1984). Our data show that there is an enrichment of full-length elements in the human genome and that many Ta elements have been faithfully replicated in their entirety and inserted into new genomic locations. Of the 399 elements examined, 119 were >6000 bp long, representing an L1 Ta size class much larger than any other (figure 2.4). By contrast, very few elements were found in the size class ranging between 3,500 and 5,500 bp, with only 22 of the 399 elements truncated to this particular size class. A bimodal distribution of the size of the elements is created, since there are a significant number of Ta L1 elements that are severely 5′ truncated and that are full-length. One hundred ninety-eight elements were extremely small, having sizes <2,000 bp, with 118 of these elements were between 25 and 1,000 bp long. The distribution is noteworthy, although the mechanism by which severely truncated and full-length L1 elements are enriched in the human genome remains to be determined. In addition, 20% (79/399) of the L1Hs elements examined are inverted at their 5′ end, which is an occurrence that is believed to be due to an event known as “twin priming” (Ostertag and Kazazian 2001), in which target primed reverse

transcription is interrupted by a second internal priming event, resulting in an inversion of the 5′ end of the newly integrated LINE. Although L1 truncation is most likely the result of the relatively low processivity of the L1 reverse transcriptase, processes like twin priming that form secondary structures in the RNA or DNA strands present at the integration site may also be associated with L1 truncation.

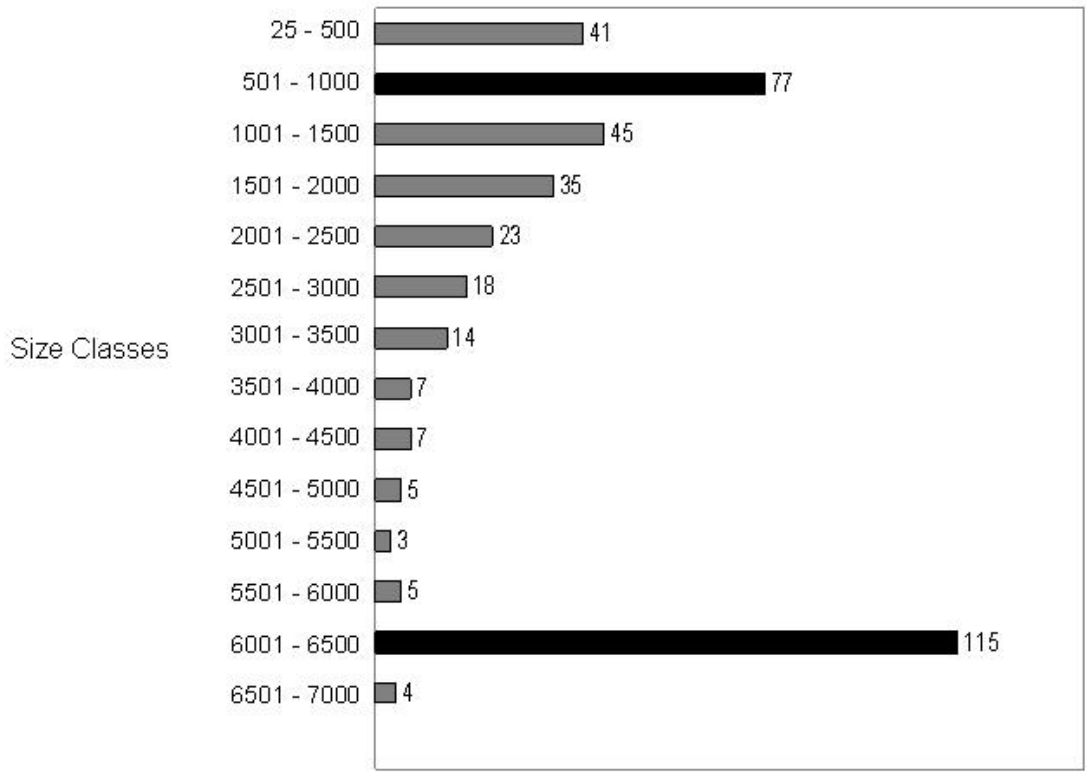

Figure 2.4

Figure 2.4 - L1 Ta element size classes (in bp), showing the size distribution of Ta L1Hs elements. Elements are grouped in 500-bp intervals ranging from <500 bp to 7,000 bp long. The two most common size intervals are shown in black.

A significant amount of sequence diversity in the 3′ tails of members of the L1HS Ta subfamily was also observed. The 3′ tails within this L1 subfamily range in size from 3 to >1,000 bp. Thirty-six percent contain AT-rich low-complexity sequence, 31% have homopolymeric A tails, 5% have simple sequence repeats with the most common repeat family TAAA, and 26% contain complex sequence that likely results from 3′ transduction events. The

diversity in the tails of the L1 elements is not surprising, since previous studies have shown an association, as well as direct evidence that mobile-element-related simple- sequence-repeat motifs mutate to form nuclei for the generation of simple sequence repeats (Economou et al. 1990; Arcot et al. 1995; Ovchinnikov et al. 2001). Three-prime transduction by L1 elements is a unique duplication event that involves retrotransposons and that has elsewhere been described in detail in L1 elements (Boeke and Pickeral 1999; Moran et al. 1999; Goodier et al. 2000). A number of 3′ transduction events that are mediated by Ta L1Hs elements were identified and suggest that these elements have transduced a total of ~8,500 bp of sequence. Computationally identification L1 element-mediated transduction event was also used in an attempt to identify putative retrotransposition-competent L1 Ta source gene. L1HS169 has a 136-bp fragment that is located outside its direct repeats and that is adjacent to its 3′ tail; this fragment is also found adjacent to the 3′ tail of L1HS28 but inside its direct repeat (figure 2.5). This suggests that L1HS28 is a daughter copy, or the progeny, of the full-length element L1HS169. In addition, AC010966 from chromosome 18 appears to be the result of a transduction event that was also generated from an L1HS169 read-through transcript. Therefore, we conclude that L1HS169 is responsible for multiple transduction events in the human genome and has produced two independent L1 integrations located on chromosomes X and 18.

Discussion

Here we report a comprehensive analysis of the dispersion and insertion polymorphism of the youngest known L1 subfamily (i.e., Ta) within the human genome. The computational approach described herein provides an efficient and high-throughput method for recovery from the human genome of Ta L1Hs elements, many of which will be polymorphic for insertion

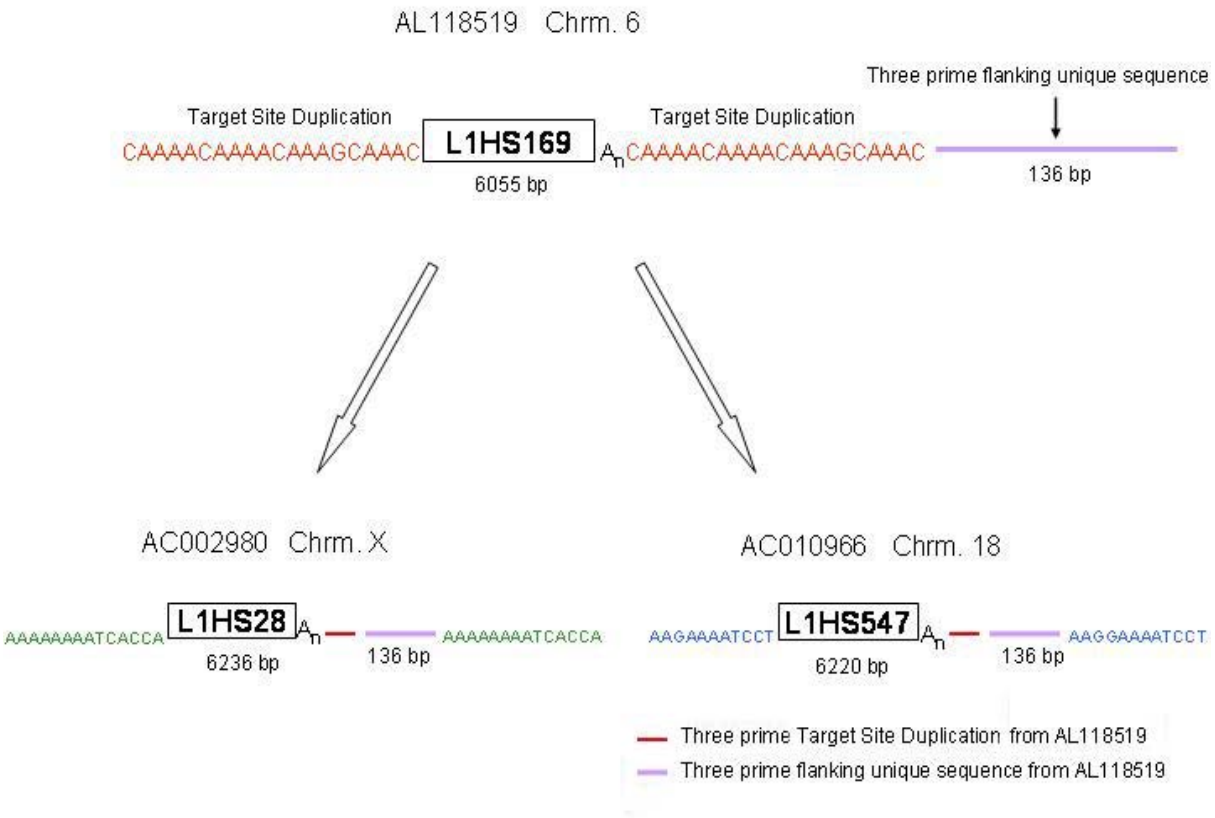

Figure 2.5

Figure 2.5 - L1HS169-mediated transduction, showing an L1Hs transduction event. L1HS169 marked by clear target site duplications is the putative source gene for L1HS28. The L1HS28 insertion contains 3′ flanking sequences identical to that of L1HS169 and unique target site duplications flanking this entire sequence, suggesting that L1HS28 was created from a read-through transcript of L1HS169 that, to give rise to L1HS28, integrated into a new location on the X. In addition, a second transduction event, L1HS547 from chromosome 18, is also flanked by unique target site duplications and was also derived from L1HS169.

presence/absence in individual humans. Individual L1 insertion polymorphisms that were identified are the products of unique insertion events within the human genome. Because each L1 element integrates into the human genome only once, individuals that share L1 insertions (and insertion polymorphisms) inherited them from a common ancestor, thereby making the L1 filled sites identical by descent. This distinguishes L1 insertion polymorphisms and other mobile element insertion polymorphisms from other types of genetic variation, including microsatellites (Nakamura et al. 1987) and RFLPs, (Botstein et al. 1980) that are not necessarily homoplasy

free. In addition, the ancestral state of an L1 insertion is known to be the absence of the L1 element. Knowledge about the ancestral state of L1 insertions facilitates the rooting of trees of population relationships by use of minimal assumptions. Therefore, the 115 new L1 insertion polymorphisms reported herein appear to have genetic properties that are similar to those of Alu insertion polymorphisms (Batzer et al. 1991; Perna et al. 1992; Batzer et al. 1994; Hammer 1994; Stoneking et al. 1997; Jorde et al. 2000), and they will serve as an additional source of identical by descent genomic variability for the study of human population relationships.

It is noteworthy that the computational identification of L1 insertion polymorphisms introduces a selection for only those elements present in the draft sequence database. As a result, elements that are not present in the database cannot be identified. This has important consequences with respect to the frequency spectrum of the elements identified. By use of this type of approach, a number of different types of L1 insertion polymorphisms are identified that vary in the frequency of the L1 insertion allele. By contrast, PCR-based display approaches provide an alternative method for the ascertainment of mobile element insertion polymorphisms from the human genome (Roy et al. 1999; Sheen et al. 2000; Ovchinnikov et al. 2001). In these approaches, polymorphic mobile elements are directly identified; however, elements that are polymorphic but have higher allele frequencies (i.e., high-frequency insertion polymorphisms) are lost in the process, since most genomes will contain at least one filled allele that contains the mobile element and would not be scored as an insertion polymorphism. Therefore, more population-specific or private mobile element insertion polymorphisms will be identified using PCR-based displays or other types of direct selection (Roy et al. 1999; Sheen et al. 2000; Ovchinnikov et al. 2001). Using our computational approach, we recovered only 14 of 49 Ta L1 elements that were elsewhere identified using PCR-based displays (Sheen et al. 2000; Ovchinnikov et al. 2001) and that had sufficient flanking unique DNA sequences for comparison

to the data set that we studied. Thus, computational and experimental ascertainment of mobile element insertion polymorphisms are quite complementary approaches for the identification of new mobile element insertion polymorphisms.

The L1 Ta subfamily can further be subdivided into Ta-0 and Ta-1, according to the nucleotides that are present at positions 5536 and 5539 within ORF 2 (Boissinot et al. 2000). Ta-0 L1 elements are believed to be evolutionarily older, and they possess a G at position 5536 and a C at position 5539. Ta-1 elements, however, have a T at position 5536 and a G at nucleotide 5539. Ta-1 elements are considered to be younger, and it is believed that all actively transposing elements in humans belong to the Ta-1 subset of L1 elements (Boissinot et al. 2000). One hundred ninety-two of the 459 Ta elements identified from the draft human genomic sequence belong to the younger Ta-1 subset, and 137 are Ta-0 subset. Another 105 of the elements either are 5′ truncated such that they terminated before these positions at 5536 and 5539 or are inverted or rearranged in the region in question. An additional 25 elements are sequence intermediates having both Ta-1 and Ta-0 diagnostic bases.

Inspection of the insertion polymorphism data for each of these Ta subsets showed that only 35% of the Ta-0 L1 elements analyzed by PCR were polymorphic, with the remaining 65% being fixed present in the human populations screened. Consistent with the idea that Ta-0 L1 elements are older, 9 of the polymorphic elements were high-frequency insertion polymorphisms, 10 were intermediate-frequency insertion polymorphisms, and only 5 were low-frequency insertion polymorphisms. None of the Ta-0 L1 elements were fixed absent or very low frequency in the populations that were analyzed. By contrast, 56% of the Ta-1 L1 elements were polymorphic with respect to presence, with 18 high-frequency, 27 intermediate-frequency, and 11 low-frequency insertion polymorphisms. In addition, we can use the non-CpG mutation density in Ta-0 and Ta-1 L1 elements to calculate the estimated age of each of the Ta-derivative

subfamilies. The non-CpG mutation density for the Ta-0 and Ta-1 L1 elements was 0.003103 and 0.002560, respectively. Using a neutral rate of evolution of 0.15% per million years (Miyamoto et al. 1987), we derive estimates of 2.07 (i.e., 0.003103/0.0015) million years and 1.71 (i.e., 0.002560/0.0015) million years from the Ta-0 and Ta-1 subsets, respectively. Although these estimates are not significantly different from each other, they do support the notion that the Ta-0 L1 elements are slightly older than the Ta-1 L1 elements, as do the differences in insertion polymorphism. In addition, they provide direct evidence that the Ta-0 and Ta-1 subsets have simultaneously amplified within the human genome.

Forty-four of the 124 full-length Ta L1Hs elements that were identified have both ORFs intact and are presumably retrotransposition-competent elements. This compares favorably with previous estimates of the number of potentially active L1 elements in the human genome (Sassaman et al. 1997). In addition, it is also important that those full-length elements that no longer have intact ORFs might have previously acted as active "source," or driver, genes for the expansion of Ta L1 elements but might have accumulated mutations over time that inactivated them. These data, as well as data from the previous studies involving the isolation and amplification of some of these full-length Ta L1 elements within tissue-culture systems, demonstrate that multiple L1 elements have expanded within the human genome in an overlapping time frame. It is interesting to compare the amplification of the L1 elements to that of the Alu SINEs within the human genome. In the case of the L1 elements, one major family (Ta) with two subdivisions (Ta-0 and Ta-1) has expanded to a copy number of ~500 elements in the past four to six million years since the divergence of humans and African apes. By contrast, the expansion of Alu elements is characterized by the amplification of at least three major lineages, or subfamilies of elements, that have collectively generated ~5000 copies (Batzer and Deininger 2002). On the basis of these copy numbers alone, it would appear that Alu elements

have been 10 times more successful than L1 elements have been with respect to duplicating themselves, within primate genomes, over the past four to six million years. However, if we make the estimate relative to the total family size of 500,000 L1 elements or 1.1 million Alu elements (Lander et al. 2001), then the relative difference is merely fivefold. This difference in amplification is also apparent across the entire expansion of these repeated DNA sequence families, since the L1 elements have expanded to only 500,000 copies in 150 million years, whereas the Alu elements have expanded to 1.1 million copies in only 65 million years.

Since Alu and L1 elements are thought to utilize the same enzymatic machinery for their mobilization, the differential amplification of both young and old Alu and L1 elements within primate genomes is quite interesting (Boeke 1997). The two different classes of repeats putatively compete for access to the same reverse transcriptase and endonuclease; thus, it is possible that Alu elements are currently more effective than the L1 elements at attracting the replication machinery within the human genome. If this competition between interspersed elements is important, then we may expect to see differential rates of L1 and Alu expansion in different nonhuman primate genomes as the elements compete for the common components involved in mobilization. Differential mobilization of SINEs and LINEs has been elsewhere reported in rodent genomes (Kim and Deininger 1996; Ostertag et al. 2000). Therefore, it would not be surprising to see something similar in nonhuman primate genomes. Alternatively, the differential amplification may reflect differences in selection against new L1 and Alu insertions within the human genome (Lander et al. 2001). Since L1 elements are typically much larger than Alu repeats, it is easy to envision that the larger insertions would be much more disruptive to the genome than the shorter Alu insertions are. This type of selection has been suggested as one potential explanation for the differential distributions of L1 elements (Boissinot et al. 2001) and of Alu and L1 elements (Lander et al. 2001; Ovchinnikov et al. 2001) throughout the human

genome. However, the argument that selection is responsible for the differential distribution of Alu sequences has recently been questioned on the grounds that older Alu sequences are not in enriched in GC-rich regions (Brookfield 2001). Further studies of the expansion of interspersed elements within the genomes of nonhuman primates will be required in order to definitively address these questions.

Our analysis of mosaic Ta L1Hs elements suggests that gene conversion alters the sequence diversity within these elements. This is not surprising, since previous studies have indicated that gene conversion plays a role in the generation of sequence diversity in Alu repeats (Maeda et al. 1988; Batzer et al. 1995; Kass et al. 1995; Roy et al. 2000; Carroll et al. 2001; Roy-Engel et al. 2002), as well as the generation of sequence diversity in L1 elements, within the genome (Hardies et al. 1986; Burton et al. 1991; Tremblay et al. 2000). Unfortunately, an accurate estimate of L1-based gene conversion is not yet possible, because primate L1 subfamily structure is not yet clearly defined. However, gene conversion appears to play a significant role in the sculpting of human genomic diversity (Ardlie et al. 2001; Frisse et al. 2001). Because of the hierarchical subfamily structure of Alu and LINEs and because of the defined pattern of ancestral mutations, these elements provide a unique opportunity for the estimation of gene conversion throughout the genome. It is also important to consider that the gene conversion between large multigene families, such as SINEs and LINEs, may occur by a mechanism that is completely different from that which occurs at other unique and low-repetition sequences within the human genome. Nevertheless, large-scale studies of orthologous sequences from the same L1 element in different human genomes will begin to quantitatively address this issue and also will provide insight into the molecular mechanism that drives the process. In addition, detailed pedigree analyses or studies of germ cell-derived L1 diversity will provide insight into the germ line rate of gene conversion between L1 elements. Clearly, L1 elements continue to have a

significant impact on human genetic diversity, through recombination, insertional mutagenesis, gene conversion, sequence transduction, and the generation of other simple-sequence-repeat motifs (Kazazian and Moran 1998; Goodier et al. 2000; Ovchinnikov et al. 2001).

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Batzer Lab, http://batzerlab.lsu.edu/

BLAST, http://www.ncbi.nlm.nih.gov/blast/

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for the DNA sequences from the common and pygmy chimpanzee orthologs of L1HS72 [accession numbers AF489459 and AF489460]; diverse DNA sequences from L1HS72 [accession numbers AF489450-AF489458]; and Ta L1 element pre-integration site sequences, namely, L1HS45 [accession numbers AF461364 and AF461365], L1HS172 [accession numbers AF461368 and AF461369], L1HS178 [accession numbers AF461370 and AF461371], L1HS284 [accession numbers AF461372 and AF461373], L1HS372 [accession numbers AF461374 and AF461375], L1HS416 [accession numbers AF461376 and AF461377], L1HS442 [accession numbers AF461378 and AF461379], L1HS443 [accession numbers AF461386 and AF461387], L1HS513 [accession numbers AF461380-AF461382], and L1HS558 [accession number AF461383])

Genetic Information Research Institute Censor Server, http://www.girinst.org/Censor_Server-Data_Entry_Forms.html

Primer3, http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi

RepeatMasker Web Server, http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) Alu repeats: a source for the genesis of primate microsatellites. Genomics 29:136-144

Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, Winchester E, Lander ES, Kruglyak L (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet 69:582-589

Ausabel FM, Brent R, Kingston ME, Moore DD, Seidman JG (1987) Current protocols in molecular biology. John Wiley & Sons, New York

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nature Reviews Genetics 3:370-379

Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ, Deininger PL (1991) Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res 19:3619-3623

Batzer MA, Rubin CM, Hellmann-Blumberg U, Alegria-Hartman M, Leeflang EP, Stern JD, Bazan HA, Shaikh TH, Deininger PL, Schmid CW (1995) Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. J Mol Biol 247:418-427

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ, et al. (1994) African origin of human-specific polymorphic Alu insertions. Proc Natl Acad Sci U S A 91:12288-12292

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499-1504

Boeke JD (1997) LINEs and Alus--the polyA connection. Nature Genetics 16:6-7

Boeke JD, Pickeral OK (1999) Retroshuffling the genomic deck. Nature 398:108-109

Boissinot S, Chevret P, Furano AV (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17:915-928

Boissinot S, Entezam A, Furano AV (2001) Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol 18:926-935

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331

Brookfield JF (2001) Selection on Alu sequences? Curr Biol 11:R900-901

Burton FH, Loeb DD, Edgell MH, Hutchison CA, 3rd (1991) L1 gene conversion or same-site transposition. Mol Biol Evol 8:609-619

Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, Myers J, Ahmad Z, Nguyen L, Sammarco M, Watkins WS, Henke J, Makalowski W, Jorde LB, Deininger PL, Batzer MA (2001) Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol 311:17-40

Cost GJ, Boeke JD (1998) Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry 37:18081-18093

Cost GJ, Golding A, Schlissel MS, Boeke JD (2001) Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res 29:573-577

Deininger PL, Batzer MA, Hutchison CA, 3rd, Edgell MH (1992) Master genes in mammalian repetitive DNA amplification. Trends Genet 8:307-311

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH, Jr. (1991) Isolation of an active human transposable element. Science 254:1805-1808

Economou EP, Bergen AW, Warren AC, Antonarakis SE (1990) The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. Proc Natl Acad Sci U S A 87:2951-2954

Eng B, Ainsworth P, Waye JS (1994) Anomalous migration of PCR products using nondenaturing polyacrylamide gel electrophoresis: the amelogenin sex-typing system. J Forensic Sci 39:1356-1359

Fanning TG, Singer MF (1987) LINE-1: a mammalian transposable element. Biochim Biophys Acta 910:203-212

Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905-916

Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci U S A 88:7396-7400

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69:831-843

Goodier JL, Ostertag EM, Kazazian HH, Jr. (2000) Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum Mol Genet 9:653-657

Grimaldi G, Skowronski J, Singer MF (1984) Defining the beginning and end of KpnI family segments. Embo J 3:1753-1759

Hammer MF (1994) A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. Mol Biol Evol 11:749-761

Hardies SC, Martin SL, Voliva CF, Hutchison CA, 3rd, Edgell MH (1986) An analysis of replacement and synonymous changes in the rodent L1 repeat family. Mol Biol Evol 3:109-125

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979-988

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 94:1872-1877

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119-121

Kass DH, Batzer MA, Deininger PL (1995) Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. Mol Cell Biol 15:19-25

Kazazian HH, Jr. (1998) Mobile elements and disease. Curr Opin Genet Dev 8:343-350

Kazazian HH, Jr. (2000) Genetics. L1 retrotransposons shape the mammalian genome. Science 289:1152-1153

Kazazian HH, Jr., Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nat Genet 19:19-24

Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332:164-166

Kim J, Deininger PL (1996) Recent amplification of rat ID sequences. J Mol Biol 261:322-327

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72:595-605

Maeda N, Wu CI, Bliska J, Reneke J (1988) Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol Biol Evol 5:1-20

Miyamoto MM, Slightom JL, Goodman M (1987) Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. Science 238:369-373

Moore JK, Haber JE (1996) Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature 383:644-646

Moran JV, DeBerardinis RJ, Kazazian HH, Jr. (1999) Exon shuffling by L1 retrotransposition. Science 283:1530-1534

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr. (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87:917-927

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato T, Taccioli G, Batzer MA, Moran JV (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31:159-165

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, White R (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616-1622

Ostertag EM, Kazazian HH, Jr. (2001) Twin priming: a proposed mechanism for the creation of inversions in l1 retrotransposition. Genome Res 11:2059-2065

Ostertag EM, Prak ET, DeBerardinis RJ, Moran JV, Kazazian HH, Jr. (2000) Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res 28:1418-1423

Ovchinnikov I, Troxel AB, Swergold GD (2001) Genomic Characterization of Recent Human LINE-1 Insertions: Evidence Supporting Random Insertion. Genome Res 11:2050-2058

Perna NT, Batzer MA, Deininger PL, Stoneking M (1992) Alu insertion polymorphism: a new type of marker for human population studies. Hum Biol 64:641-648

Prak ET, Kazazian HH, Jr. (2000) Mobile elements and the human genome. Nat Rev Genet 1:134-144

Rothbarth K, Hunziker A, Stammer H, Werner D (2001) Promoter of the gene encoding the 16 kDa DNA-binding and apoptosis-inducing C1D protein. Biochim Biophys Acta 1518:271-275

Roy AM, Carroll ML, Kass DH, Nguyen SV, Salem AH, Batzer MA, Deininger PL (1999) Recently integrated human Alu repeats: finding needles in the haystack. Genetica 107:149-161

Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL (2000) Potential gene conversion and source genes for recently integrated Alu elements. Genome Res 10:1485-1495

Roy-Engel AM, Carroll ML, El-Sawy M, Salem AE, Garber RK, Nguyen SV, Deininger PL, Batzer MA (2002) Non-traditional Alu evolution and primate genomic diversity. Journal of Molecular Biology 316:1033-1040

Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, Batzer MA, Deininger PL (2001) Alu insertion polymorphisms for the study of human genomic diversity. Genetics 159:279-290

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463-5467

Santos FR, Pandya A, Kayser M, Mitchell RJ, Liu A, Singh L, Destro-Bisol G, Novelletto A, Qamar R, Mehdi SQ, Adhikari R, de Knijff P, Tyler-Smith C (2000) A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. Hum Mol Genet 9:421-430

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH, Jr. (1997) Many human L1 elements are capable of retrotransposition. Nat Genet 16:37-43

Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD (2000) Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res 10:1496-1508

Skowronski J, Fanning TG, Singer MF (1988) Unit-length line-1 transcripts in human teratocarcinoma cells. Mol Cell Biol 8:1385-1397

Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9:657-663

Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246:401-417

Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA (1997) Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res 7:1061-1071

Teng SC, Kim B, Gabriel A (1996) Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. Nature 383:641-644

Tremblay A, Jasin M, Chartrand P (2000) A double-strand break in a chromosomal LINE element can be repaired by gene conversion with various endogenous LINE elements in mouse cells. Mol Cell Biol 20:54-60

Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R (1998) Apolipoprotein(a) gene enhancer resides within a LINE element. J Biol Chem 273:891-897

CHAPTER THREE:

LINE-1 PRE Ta ELEMENTS IN THE HUMAN GENOME*

***Reprinted by permission of The Journal of Molecular Biology**

Introduction

Computational analysis of the draft sequence of the human genome indicates that repetitive sequences comprise 45-50% of the human genome mass, 17% of which consists of Long INterspersed Elements (LINE-1s or L1s) (Smit 1999; Prak and Kazazian 2000; Lander et al. 2001). L1 elements are restricted to mammals, having expanded as a repeated DNA sequence family over the last 150 million years (Smit et al. 1995). Full-length L1 elements are approximately 6 kilobases (kb) long and propagate via an RNA intermediate in a process known as retrotransposition. L1 retrotransposition likely occurs by a mechanism termed target primed reverse transcription (TPRT) (Luan et al. 1993). This mechanism of mobilization provides two useful landmarks for the identification of young L1 inserts: an endonuclease related cleavage site (Jurka 1997; Cost and Boeke 1998; Cost et al. 2001) and direct repeats or target site duplications flanking newly integrated elements (Fanning and Singer 1987).

L1 retrotransposons have had a significant impact on the human genome through a variety of different mechanisms. *De novo* insertions disrupting open reading frames and splice sites have resulted in a number of human diseases (Kazazian et al. 1988; Kazazian 1998; Deininger and Batzer 1999), new L1 integrations have been shown to have the potential to alter gene expression (Yang et al. 1998; Rothbarth et al. 2001), and once in the genome L1 elements provide regions of sequence identity blanketing the genome, that can be exploited during recombination (Fitch et al. 1991). L1 elements also generate sequence duplications by transducing adjacent genomic sequences at their 3' end, thereby "shuffling" genomic sequence (Boeke and Pickeral 1999; Moran et al. 1999; Goodier et al. 2000). More recently, it has been suggested that L1 elements have paradoxical roles in genomic stability by serving both as molecular band aids, repairing double stranded breaks in mammalian cells and as suspects for the generation of genomic deletions (Gilbert et al. 2002; Kazazian and Goodier 2002; Symer et al.

2002). Thus, L1 elements exert a significant influence on the architecture of the human genome and provide dynamic units capable of ongoing change.

As a result of the limited amplification potential of the diverse L1 gene family, a series of discrete L1 subfamilies exists within the human genome (Deininger et al. 1992; Smit et al. 1995). L1 elements have expanded at different times during mammalian evolution, producing subfamilies of various ages (Deininger et al. 1992; Smit et al. 1995). Depending on the amplification period of the L1 subfamily, some L1 elements may be unique to a single phylogenetic lineage, species, or even a single population. Such is the case with the L1Hs (Human specific) Ta (transcribed, subset a) (Skowronski et al. 1988) subfamily, which has been shown to be present only in the human species (Myers et al. 2002).

Even though there are approximately 500,000 L1 elements in the human genome only a limited subset of 30-60 L1 elements appears to be capable of retrotransposition (Moran et al. 1996; Sassaman et al. 1997). *De novo* L1 insertions resulting in human disease are largely the product of L1Hs Ta integrations, which have been shown to be the youngest, most active L1 subfamily found in the human genome (Boissinot et al. 2000; Sheen et al. 2000; Myers et al. 2002). However, at least one L1 insert (JH-28) in exon 14 of the factor VIII gene resulting in hemophilia A was the result of a preTa insertion, providing the first proof that preTa L1 elements are also currently capable of retrotransposition (Kazazian et al. 1988). Previous studies have shown that some members of the preTa L1 subfamily have inserted so recently in the human genome that they are polymorphic with respect to insertion presence/absence (Boissinot et al. 2000; Ovchinnikov et al. 2002), all of which makes preTa L1 elements a likely source of identical-by-descent mobile element based variation for the study of human population genetics.

Members of the L1 preTa subfamily share a common three base pair diagnostic sequence within the 3' untranslated region (UTR), which separates them from the other L1 subfamilies. As

the name suggests, the preTa L1 subfamily is believed to predate the amplification of the L1Hs Ta subfamily in the human lineage. However, the phylogenetic origin and level of human genetic diversity associated with preTa L1 elements remains largely undefined. The following work provides a comprehensive analysis of the preTa L1 subfamily from the draft sequence of the human genome.

Results

L1 preTa Subfamily Copy Number

To identify recently integrated preTa subfamily L1 elements from the human genome, the draft sequence of the human genome (database version: BLASTN 2.2.1 [Apr-13-2001]) was searched using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) with an oligonucleotide sequence that is complementary to a highly conserved motif in the 3' untranslated region (UTR) of preTa L1 elements. This 19 base pair (bp) query sequence (CCTAATGCTAGATGACACG) includes the preTa subfamily-specific diagnostic mutation "ACG" at its 3' end (position 5930-5932 relative to LRE-1) (Dombroski et al. 1991). Three hundred sixty-two unique preTa L1 elements were identifiedfrom 2.868 $x10^9$ bp of available human draft sequence. Extrapolating this number to the actual size of the human genome (3.162 $x10^9$ bp), provides an estimate that this subfamily contains about 400 elements. Taken with the estimate from the L1Hs Ta data (Myers et al. 2002), suggests that there are over 900 human specific LINE-1 elements in the human genome. Of the 362 preTa L1 elements retrieved, 6 resided at the end of sequence contigs and were not amenable to additional analysis. One hundred five (29%) of the 356 (362-6) remaining elements were essentially full length, and 251 were truncated to variable lengths. Alignment and sequence analysis of the full-length elements revealed that 29 contained two intact open reading frames and therefore may be capable of

retrotransposition. The complete data set is available on the Batzer Laboratory of Comparative Genomics website (http://batzerlab.lsu.edu) under publications.

Estimated Subfamily Age

The average ages of L1 elements can be determined by the level of sequence divergence from the subfamily consensus sequence using a neutral mutation rate for primate non-coding sequence of 0.15% per million years (Miyamoto et al. 1987). The mutation rate is known to be about 10 times greater for CpG bases as compared to non-CpG bases, as result of the spontaneous deamination of 5-methyl cytosine (Bird 1980). Thus, two age estimates based upon CpG and non-CpG mutations can be calculated for the preTa subfamily of L1 elements. A total of 74,048 bases from the 3' UTR of 356 preTa L1 elements was analyzed. Three hundred sixty-one total nucleotide substitutions were observed. Of these, 303 were classified as non-CpG mutations against the backdrop of 71,912 total non-CpG bases, producing a non-CpG mutation density of 0.004213 (303/71,912). Based upon the non-CpG mutation density and a neutral rate of evolution (0.004213/0.0015) the average age of the L1 preTa LINE-1 elements was 2.81 million years old. A total of 58 CpG mutations out of 2,136 total CpG nucleotides was found across the same 356 LINE elements, yielding a CpG based mutation density of 0.027154 (58/2,136). With the expectation that the CpG mutation rate is about 10 fold higher than the non-CpG mutation rate, the approximate age of the L1 preTa subfamily using the CpG mutation density is 1.86 million years old. These estimates are in good agreement with one another and taken together, these estimates produce an average age of 2.34 million years old, which is in good agreement with the idea that the preTa L1 subfamily is evolutionarily older than the L1Hs Ta subfamily (estimated average age 1.99 million years) (Boissinot et al. 2000; Myers et al. 2002). In addition the average age estimates reported here provide a relative time frame by

which to compare L1 retrotransposition activity, and should not be confused with the age of origin.

Similar to the L1Hs Ta subfamily, the preTa L1 subfamily can also be grouped into two subgroups, ACG/A and ACG/G, based on an "A" or "G" base at position 6015 relative to L1.2 (Accession number M80343). In order to determine the relative ages of each subgroup, we analyzed the level of sequence divergence in each subgroup. The ACG/A subgroup contained 127 total nucleotide substitutions with 98 of these classified as non-CpG mutations against the backdrop of 20,402 total non-CpG bases. This yields a non-CpG mutation density of 0.004803 (98/20,402) and produces an estimated age of 3.20 million years old. Twenty nine of 127 total mutations were classified as CpG mutations against a backdrop of 606 CpG total bases, which yields a CpG mutation density of 0.047855 (29/606) producing an estimated age of 3.28 million years. The ACG/G subgroup contained 221 total nucleotide substitutions with 191 of these classified as non-CpG mutations against the backdrop of 51,106 total non-CpG bases, which yields a non-CpG mutation density of 0.003737 (191/51,106), producing an estimated age of 2.49 million years old. Thirty of 121 total mutations were classified as CpG mutations against a backdrop of 1518 CpG total bases, which yields a CpG mutation density of 0.019763 (30/1518) producing an estimated age of 1.35 million years. We calculated the average age of each subgroup as 1.92 and 3.24 million years for the ACG/G and ACG/A respectively. Although it is likely that the L1Hs Ta subfamily is derived from one of the preTa L1 subsets based on the estimated ages of these L1 subfamilies, the transition intermediates between preTa and Ta subfamilies are not clear.

Features of L1 preTa Integration Sites

One hallmark of L1 integration is the generation of target site duplications flanking newly integrated elements. Two thousand bases of flanking sequence on each side of the

element were searched for target site duplications. Clear target site duplications are considered to be target site duplications at least 10 bases in length. Of the 356 elements analyzed, we were able to identify clear target site duplications for 252 elements. We then determined the integration sites for these 252 preTa L1 insertions with clear target duplications. A complete list of L1 integration sites is shown in Table 3.1, and further supports the notion that some integration sites are more common than others (Feng et al. 1996; Jurka 1997; Cost and Boeke 1998).

Table 3.1 - PreTa L1 integration sites.

preTa L1 Integration Sites	Number
TTTT/A	60
TCTT/A	37
CTTT/A	20
TTTA/A	18
TTTC/A	18
TTTT/G	16
TTCT/A	14
TCTT/G	7
CTTT/G	5
ATTT/A	5
CTTT/C	5
TTTT/C	4
TGTT/A	3
TATT/A	3
TATT/G	3
TCTT/C	2
TTTC/C	2
TCTC/A	2
GTTT/A	2
ATTT/C	2
GCTT/T,TTTT/T,TTTG/A,TTTC/T,TTTC/T,TTGT/G, TTAT/A,TGAT/G,TCTT/T,TCAT/A,TATC/A,TATA/T, TAAA/C,GCTT/A,CCTT/A,CATT/G,CATT/A,ACTT/G, ACTT/A,ACTA/C,ACCT/A,ACAC/T,ACAA/A,AAAA/A	1 each

A large number of preTa L1 elements had no observable target duplication sites. One possible explanation for this observation is that these elements have relatively short target site

duplications. Alternatively, these elements may represent forward gene conversion events of older pre-existing L1 elements that by mutation, have rendered their target site duplications unrecognizable. Some of these events may also represent integrations that have occurred independent of endonuclease cleavage, that has previously been proposed as a mechanism for the repair of doubled stranded breaks in DNA (Moore and Haber 1996; Teng et al. 1996; Morrish et al. 2002).

To further characterize the preTa L1 insertions, we determined the DNA base content for sequence blocks 1 and 2 kb flanking all preTa L1 insertion sites with target site duplications of at least 10 bp. Flanking sequence was then grouped according to GC content with only data for the 1 kb sequence blocks shown in Figure 3.1. Our data suggests that preTa L1 elements integrate preferentially in genomic regions with GC content less than 36%, but are present in genomic regions with GC content as low as 26% and as high as 52%. A similar insertion site preference was observed for 2 kb sequence blocks as well as for the previously reported L1 Ta subfamily (Myers et al. 2002) and other L1 subfamilies (Szak et al. 2002). In addition, we also analyzed preTa L1 elements inserted in repetitive sequences and grouped them according to the repeat family in which they reside (Figure 3.2). This analysis showed that preTa L1 elements insert most frequently in other L1 elements, which is expected both because L1 sequences occupy a large percentage of the human genome and because L1 elements are less GC rich relative to other mobile element families, such as Alu elements, making them more susceptible to subsequent L1 integrations. Lastly, preTa L1 containing regions were analyzed to determine the distance from the integration to the nearest gene. Twelve preTa L1 elements reside within 25 kb of novel or known genes as denoted by GenBank annotation, including one full length preTa element, L1AD242, which inserted into intron 23-24 of the retinoblastoma susceptibility protein 1 gene and accounts for 6072 bp of the 7988 bp intron.

Sequence Diversity

PreTa L1 sequence diversity is also created by variable 5′ truncation with some of the elements in the human genome only a few hundred base pairs in length, whereas some full-length elements are over 6000 base pairs. This phenomenon is classically attributed to the lack of processivity of the reverse transcriptase enzyme in the creation of the L1 cDNA. The point of truncation is traditionally believed to occur as a function of length, where shorter inserts are more likely to occur in the human genome than longer elements (Grimaldi et al. 1984). Our data show that there is an enrichment of full-length elements in the human genome, and like the Ta L1 elements many preTa L1 elements have been faithfully replicated in their entirety and inserted into new genomic locations. Of the 356 elements examined (362 total minus 6 elements located at the end of sequencing contigs), 97 were over 6000 base pairs long, representing a much larger preTa L1 size class than any other size class (Figure 3.3). By contrast, very few elements were found in the size ranges between 4000 and 5500 bases, with only 14 of the 356 elements truncated to this particular size range. A bimodal distribution in the size of the elements is created since there are a significant number of preTa L1 elements that are severely 5 prime truncated and those that are full-length with the average preTa element length of roughly 2700 bp and the median preTa element length of roughly 1600 bp. One hundred ninety-six elements were small with sizes less than 2000 bp, with 125 of these only 50-1000 bases in length. In addition 28% (100/356) of the preTa L1 elements examined were inverted at their 5 prime end, which is believed to occur by an event known as twin priming where target primed reverse transcription is interrupted by a second internal priming event, resulting in an inversion of the 5 prime end of the newly integrated LINE element (Ostertag and Kazazian 2001). Although L1 truncation is most likely the result of the relatively low processivity of the L1 reverse transcriptase, processes that form secondary structures in the RNA or DNA strands

present at the integration site, like twin priming, may also be associated with L1 truncation. One expectation of this model is that a common truncation point should exist for L1 preTa elements. However, from our data we were not able to identify any common truncation points. Similar to other L1 elements, preTa L1 elements exhibit a significant amount of sequence diversity in the 3 prime tails. In general, the 3 prime tails found in this L1 subfamily range in size from 4 to over 1600 bp in length. Sixty-four percent contain AT rich low complexity sequence, 13% have homopolymeric A tails with an average tail length of 15 bp, 6% have simple sequence repeats with the most common repeat family $TAAA_n$, and 17% contain complex sequence likely resulting from 3 prime transduction events. Three-prime transduction by L1 elements is a unique duplication event that occurs when an L1 sequence is transcribed along with genomic sequence at its 3 prime end. This sequence then integrates at a different genomic location resulting in duplication of the source L1 sequence and the 3 prime genomic sequence flanked by target site duplications (Boeke and Pickeral 1999; Moran et al. 1999; Goodier et al. 2000). We have

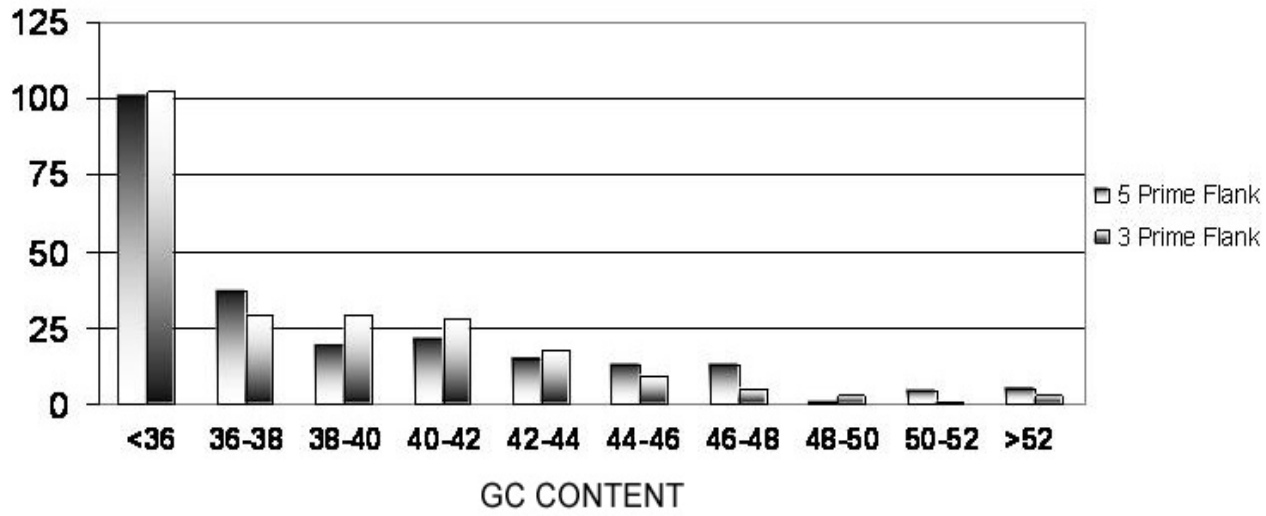

Figure 3.1

Figure 3.1 – Analysis of preTa L1 pre-integration sites. GC content was calculated for L1 insertion flanking sequences of 1 and 2 kb. The 1 kb results are shown here.

identified fifty 3 prime transduction events mediated by preTa L1 elements and believe that these elements have transduced approximately 10,400 total bases of sequence with one transduction event responsible for duplicating a region over 1600 bp. The diversity observed in the tails of the L1 elements is not surprising since previous studies have shown an association as well as direct evidence that simple sequence repeat motifs present in the 3 prime tail of mobile elements can mutate serving as nuclei for the generation of simple sequence repeats (Economou et al. 1990; Arcot et al. 1995; Ovchinnikov et al. 2001). A complete list of the preTa elements involved in transduction events is located at the Batzer Laboratory of Comparative Genomics website under publications (http://batzerlab.lsu.edu).

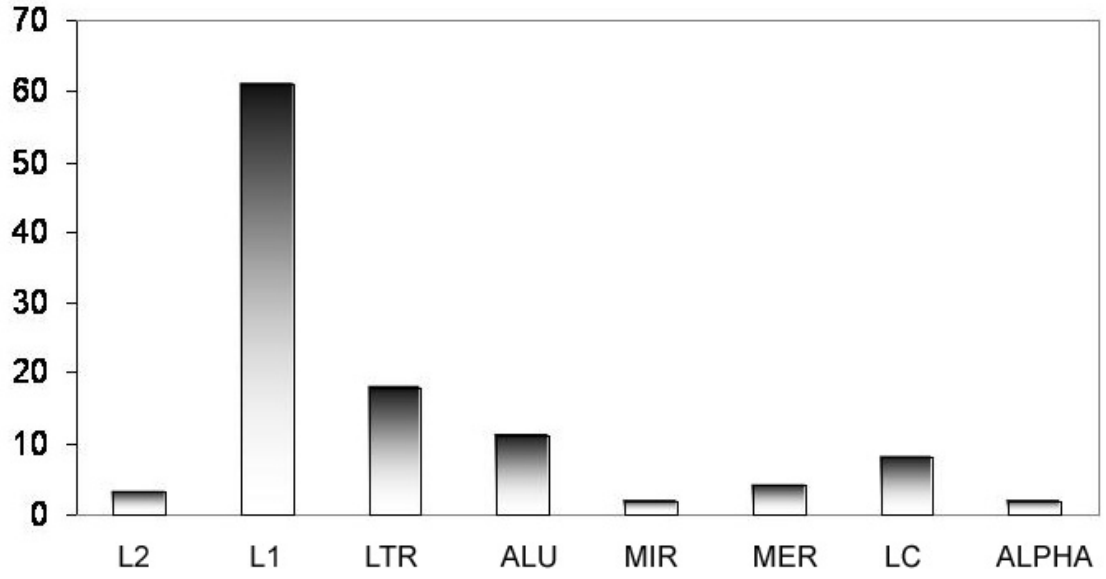

Figure 3.2

Figure 3.2 - PreTa L1 integrations within other repetitive elements. PreTa L1 insertions within mobile elements were grouped according to the element in which they inserted. Mobile elements categories include LINE-2 (L2), LINE-1 (L1), Long Terminal Repeats (LTR), Alu (ALU), Mammalian-wide Interspersed Repeats (MIR), Medium Reiteration Frequency Sequences (MER), Low Complexity Sequence (LC), Alphoid Satellite Repeats (ALPHA).

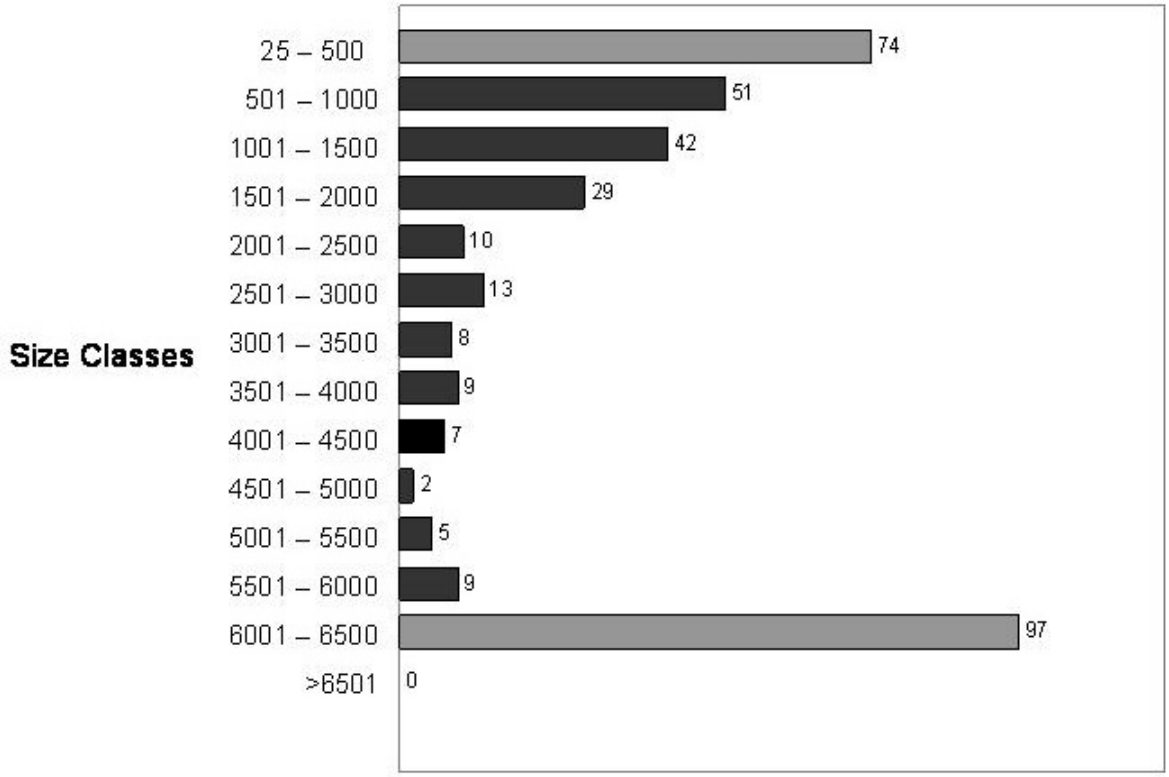

Figure 3.3

Figure 3.3 - PreTa L1 element genomic size distribution. The following schematic shows the size distribution of preTa L1 elements. Elements are grouped in 500 bp intervals ranging from 25 bp in length to >6501 bp in length. The two most common size intervals are denoted in gray.

L1 Associated Human Genomic Diversity

Of the 362 preTa L1 elements isolated *in silico*, 102 of the elements were inserted into other repetitive regions of the genome such that flanking unique sequence PCR primers could not be designed. Six additional elements resided at the end of sequencing contigs in GenBank and lacked unique flanking sequence information making PCR primer design in this region impossible. The remaining 254 were analyzed using a subfamily specific polymerase chain reaction assay and flanking unique sequence primers as previously described (Sheen et al. 2000) (summarized in Table 3.2). Three elements out of 254, produced inconclusive PCR results because of the amplification of paralogous genomic sequences as described previously (Batzer et

al. 1991). Nine elements produced non-specific PCR results, and were excluded from further analysis. Another nine elements produced subfamily-specific PCR products in all human samples tested, but did not produce pre-integration site in both human and non-human primate genomes. This may be the result of some type of large deletion event that occurred in the human genome and not in the genome of non-human primates making the non-human primate pre-integration site much larger than expected and not detectable by our assay as reported previously (Myers et al. 2002). Alternatively this could also be the result of mutations in the oligonucleotide hybridization sites rendering them ineffective for PCR. In addition, we identified 36 preTa L1 elements that mapped to the X chromosome and eight that mapped to the Y chromosome, all of which were fixed present in the individuals tested (APPENDIX B Supplementary Data Table 3). The human genomic diversity associated with the autosomal preTa L1 elements is shown in Supplementary Data Table 3 and Supplementary Data Table 4 (APPENDIX B).

Table 3.2 - Summary of preTa L1 analysis.

LOCI ANALYZED BY PCR	254
Fixed present	200
High frequency insertion polymorphisms	11
Intermediate frequency insertion polymorphisms	22
Low frequency insertion polymorphisms	0
Total preTa insertion polymorphisms	33
Inserted in paralogous sequences	3
No pre-integration site amplified in primates	9
No PCR results	9
LOCI NOT ANALYZED BY PCR	
L1 elements inserted in other repeats	102
End of contig	6
Total preTa L1 elements analyzed	362

Two hundred thirty three (254-9-9-3) preTa L1 elements produced unambiguous results when analyzed by a two-step PCR assay across 80 individuals from four geographically diverse human populations with 33 (14%) being polymorphic with respect to insertion presence/absence (APPENDIX B Supplementary Data Table 3 and 4). Examples of human genomic diversity associated with preTa L1 insertion polymorphisms are shown in Figure 3.4A and 3.4B. Eleven of the preTa L1 elements were high frequency insertion polymorphisms with L1 element allele frequencies greater than 0.70, so that most of the individuals were homozygous (+/+) for the presence of the LINE element. Twenty-two of the polymorphic elements were intermediate frequency, with a LINE element allele frequency greater than 0.30 but less than 0.70 across the diverse human populations sampled. None of the L1 preTa elements tested had insertion allele frequencies less than 0.30. One possible explanation for the absence of low frequency preTa insertion polymorphisms would be that the preTa subfamily has largely undergone retrotranspositional quiescence and is no longer generating new copies. As a result, the number of low frequency preTa insertion polymorphisms in the human genome would be limited. It is also possible that the newly integrated preTa L1 elements are removed from the human genome as a result of negative selection. However, we consider the former explanation more likely based upon the three-fold higher levels of insertion polymorphism in the Ta subfamily as compared to the preTa subfamily (45% vs. 14%) as well as the previously reported frequency distribution of Ta L1 insertion polymorphisms in the human genome (Myers et al. 2002).

Two hundred preTa L1 elements were fixed present in the human genome. These elements are likely to be slightly older than their polymorphic counterparts, having inserted into the human genome prior to the radiation of humans from Africa. Overall, the unbiased heterozygosity values across all of the L1 elements subjected to PCR analysis were similar cross the four populations with values of 0.306 in African Americans, 0.243 in Asians, 0.252 in

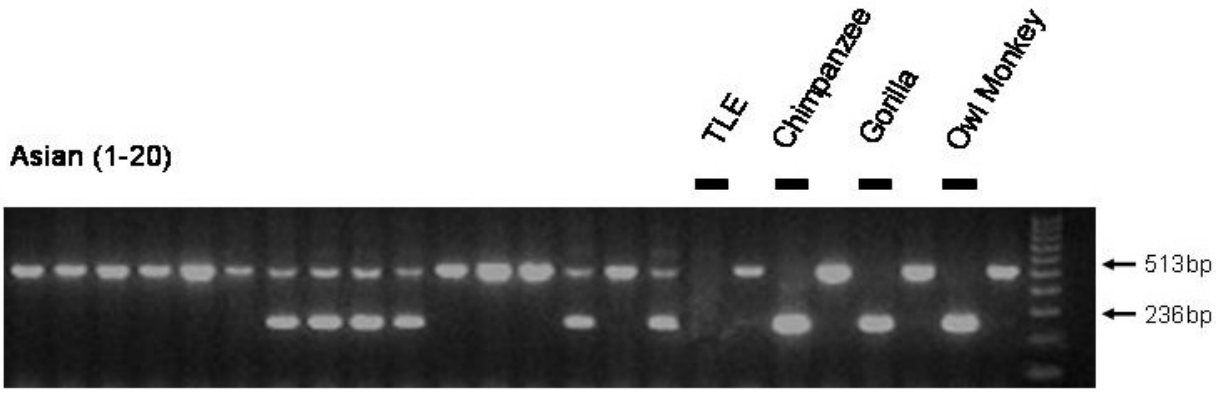

Figure 3.4A

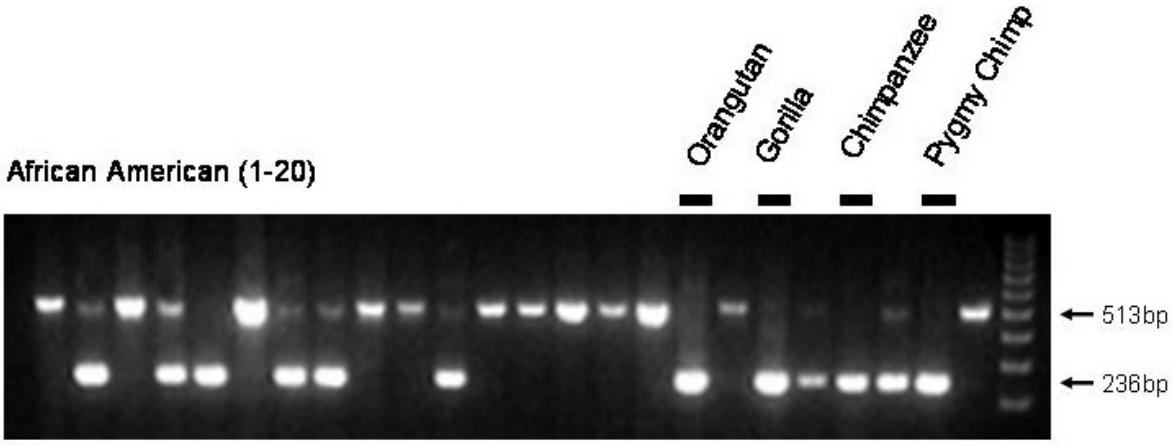

Figure 3.4B

Figure 3.4 - PreTa L1 insertion polymorphisms. This figure is an agarose gel chromatograph of the PCR products from a survey of the human genomic variation associated with L1AD125. Amplification of the pre-integration site of this locus generates a 236 bp PCR product. Amplification of a filled site generates a 513 bp product (using flanking unique sequence primers). In this survey of human genomic variation 20 individuals from each of four diverse populations were assayed for the presence or absence of the L1 element, with Asian samples shown in Figure 3.4A and African Americans shown in Figure 3.4B. The control samples are denoted by the black lines and were TLE buffer (10 mM Tris-HCl: 0.1 mM EDTA), common chimpanzee, pygmy chimpanzee, gorilla, orangutan and owl monkey DNA templates. In addition, this particular L1 element was absent from the genomes of non-human primates.

Europeans, and 0.269 in South Americans, with the African American population being the most diverse with respect to preTa L1 alleles (APPENDIX B Supplementary Data Table 4). However, several of the polymorphic elements individually exhibited unbiased heterozygosity values that approached 0.5, the theoretical maximum for bi-allelic loci.

To determine if the LINE insertion polymorphisms were in Hardy-Weinberg Equilibrium (HWE), expected genotype frequencies were compared with observed genotype frequency using chi-square tests for goodness of fit. A total of 132 chi-square tests for goodness of fit are theoretically possible. However, 28 of the comparisons involved populations that were monomorphic for the presence of the L1 insertion leaving 104 possible tests. A total of 23 deviations from Hardy-Weinberg expectations were observed in the comparisons. Eighteen of the deviations were the result of low expected genotype frequencies. Of the remaining five tests that deviated from HWE, none clustered by population or locus. This deviation is not surprising since a total of 5.15 deviations from HWE would be expected by chance alone at the 5% significance level. One short coming of this method is its inability to deal with low expected genotype frequencies. To further test these polymorphisms for HWE, we performed an exact test for Hardy-Weinberg proportions using the Markov chain test available in the Arlequin program (Guo and Thompson 1992) , which is not hindered by low expected frequencies. The exact test showed that none of the 104 comparisons deviated from HWE proportions at the one percent level. Therefore we conclude that the newly identified L1 insertion polymorphisms do not significantly depart from HWE.

Discussion

Here we report a comprehensive analysis of the dispersion and insertion polymorphism associated with the preTa L1 subfamily within the human genome. We estimate that there are approximately 900 lineage specific L1 elements present in the entire human genome. In

addition, given the median size for preTa and Ta L1 elements (~1600 bp) and a conservative copy number estimate of 900 elements, we estimate that human lineage-specific L1 retrotransposition has been responsible for increasing the size of the human genome by roughly 1.4 million bases.

The level of sequence diversity, estimated age, and the reduction of human genomic variation associated with this L1 subfamily relative to the Ta L1 subfamily provide strong evidence suggesting the expansion of preTa L1 elements began prior to the expansion of the Ta L1 subfamily that has been analyzed in detail previously (Boissinot et al. 2000; Myers et al. 2002). However, the expansion of preTa L1 elements also appears to have occurred over a time frame that predated the radiation of humans from Africa and continued until very recently, in fact it may still be occurring at a very low level within the human lineage. Thus, we conclude that the expansion of preTa and Ta L1 elements occurred in an overlapping time frame in the human lineage. The reason(s) for the relative retrotranspositional quiescence of preTa elements remain unknown. However, they may relate to alterations in the ORF2 protein of the preTa elements, decreased transcription from the preTa "source" elements or a decrease in the ability of the elements to undergo target primed reverse transcription (Moran 1999). Further studies using *in vitro* systems to measure retrotransposition (Moran et al. 1996) will be required to definitively address this question.

Sequence analysis of the preTa L1 insertions suggest that they have a slight preference for integrating into regions of the genome with low GC content. This observation is contradictory to that previously reported (Ovchinnikov et al. 2001), but is in agreement with results obtained by The International Human Genome Sequencing Consortium (Lander et al. 2001). The reason for this integration site preference is unclear, but may result from a subtle sequence preference of the preTa encoded endonuclease. Alternatively, this observation may

reflect limitations on L1 preTa insertion events imposed by chromatin organization. However, it is likely that both factors, as well as others not mentioned here, are important in determining where in the human genome young L1 elements will integrate. It is also interesting to note that some preTa L1 insertions have occurred adjacent to known genes. The persistence of these newly integrated preTa L1 elements in these regions of the human genome is most likely indicative that they have had no negative effects with respect to the function of these genes.

Twenty-nine of the essentially 105 full length L1 preTa elements identified have both open reading frames intact and are putatively retrotransposition competent elements. The data collected from the L1 preTa subfamily along with the L1Hs Ta subfamily (44 elements) yields a computational estimate of 73 active L1 elements within the genome that is comparable to previous estimates of the number of potentially active L1 elements in the human genome (Sassaman et al. 1997). Collectively, these data suggest L1 elements from multiple subfamilies may still be capable of retrotransposition within the human lineage. In addition, it is also important to mention that those full-length elements that no longer have intact open reading frames could have previously served as active "source" or driver genes for the expansion of pre Ta L1 elements, but have accumulated mutations over time that subsequently inactivated them.

The computational identification approach described here provides an efficient and high-throughput method for recovering preTa L1 elements from the human genome, some of which are polymorphic for insertion presence/absence in individual human genomes. Individual L1 insertion polymorphisms identified, similar to other mobile element insertion polymorphisms, are the products of unique insertion events within the human genome. Because each L1 element integrates only once into the human genome, individuals that share L1 insertions (and insertion polymorphisms) inherited them from a common ancestor, making the L1 filled sites identical by descent (Sheen et al. 2000; Myers et al. 2002). This distinguishes L1 insertion polymorphisms

from other types of genetic variation that may not be derived from a single ancestral allele including microsatellites (Nakamura et al. 1987) and restriction fragment length polymorphisms (Botstein et al. 1980; Nakamura et al. 1987). In addition, the ancestral state of an L1 insertion is known to be the absence of the L1 element. Therefore the thirty-three new L1 insertion polymorphisms reported here appear to have genetic properties similar to the previously identified Alu (Batzer et al. 1991; Perna et al. 1992; Batzer et al. 1994; Hammer 1994; Stoneking et al. 1997; Jorde et al. 2000) and L1 (Boissinot et al. 2000; Sheen et al. 2000; Myers et al. 2002) insertion polymorphisms and provide a unique form of genetic variation present in the human population that will serve as an additional source of identical by descent genomic variability for the study of human population relationships.

Materials and Methods

Cell lines and DNA samples

The cell lines used to isolate primate DNA samples were as follows: human (*Homo sapiens*) HeLa (ATCC CCL2), common chimpanzee (*Pan troglodytes*) Wes (ATCC CRL1609), pygmy chimpanzee (*Pan paniscus*) Coriell Cell Repository Number AG05253, gorilla (*Gorilla gorilla*) Lowland Gorilla (Coriell Cell Repository Number AG05251B), green monkey (*Cercopithecus aethiops*) ATCC CCL70, owl monkey (*Aotus trivirgatus*) OWK (OWKidney) ATCC CRL 1556, and Orangutan (*Pongo pygmaeus*) (Coriell Primate Panel PRP00001 Cell Repository Number NG12256) . Cell lines were maintained as directed by the source and DNA isolations were performed using Wizard genomic DNA purification (Promega). Human DNA samples from the European, African American, and Asian population groups were isolated from peripheral blood lymphocytes (Ausabel et al. 1987) available from previous studies (Stoneking et al. 1997). South American Human DNA was obtained form Coriell Human Variation Panels HD17 and HD18.

Computational Analyses

The draft sequence of the human genome was screened using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) available at the National Center of Biotechnology Information Genomic Blast page (http://www.ncbi.nlm.nih.gov/BLAST/). A 19 base pair oligonucleotide, 5'-CCTAATGCTAGATGACACG-3' that is diagnostic for the preTa subfamily was used to query the Human Genome database with the following the optional parameters: filter none; advanced options -e 0.1, -v 600, -b 600. Copy number estimates were determined from BLAST search results. Sequences containing exact matches were subjected to additional analysis as outlined below.

A sequence region of 9000-10000 bases, including the match and 1000-2000 bases of flanking unique sequence were annotated using RepeatMasker version 7/16/00 from the University of Washington Genome Center Server (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) or Censor from the Genetic Information Research Institute (http://www.girinst.org/Censor_Server-Data_Entry_Forms.html) (Jurka et al. 1996). These programs annotate repeat sequence content and were used to confirm the presence of preTa L1 elements and regions of unique sequence flanking the elements. PCR primers flanking each L1 element were designed using Primer3 software available at the Whitehead Institute for Biomedical Research (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgiand) and were complementary to the unique sequence regions flanking each L1 element. The resultant primers were screened with standard nucleotide-nucleotide BLAST [blastn] against the non-redundant (nr) and high-throughput (htgs) sequence databases to ensure they resided in unique DNA sequences. Primers residing in repetitive sequence regions were discarded and new primers designed if possible. A complete list of all the L1 elements identified using this approach is available from our website (http://batzerlab.lsu.edu). Individual L1 DNA sequences

were aligned using MegAlign with the ClustalW algorithm and the default settings (DNAstar version 5.0 for Windows) followed by manual refinement.

PCR Amplification

PCR amplification of 255 individual L1 elements was carried out in 25 µl reactions containing 20-100 ng of DNA of template DNA, 40 pM of each oligonucleotide primer (APPENDIX B Supplementary Data Table 3), 200 µM dNTPs, in 50 mM KCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl (pH 8.4) and Taq DNA polymerase (1.25 Units). Each sample was subjected to the following amplification for 32 cycles: an initial denaturation of 150 seconds at 94 °C, one minute denaturation at 94 °C, one minute at the annealing temperature (specific for each locus), and an extension at 72 °C for one minute. Following the cycles a final extension was performed at 72 °C for ten minutes. For analysis, 20 µl of each sample was fractionated on a 2% agarose gel with 0.05 µg/ml ethidium bromide. PCR products were directly visualized using UV fluorescence. The human genomic diversity associated with each L1 preTa element was determined by the amplification of 20 individuals from each of four geographically distinct populations (African American, Asian, European, and South American) for a total of 160 chromosomes.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) Alu repeats: a source for the genesis of primate microsatellites. Genomics 29:136-144

Ausabel FM, Brent R, Kingston ME, Moore DD, Seidman JG (1987) Current protocols in molecular biology. John Wiley & Sons, New York

Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ, Deininger PL (1991) Amplification dynamics of human-specific (HS) Alu family members. Nucleic Acids Res 19:3619-3623

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ, et al. (1994) African origin of human-specific polymorphic Alu insertions. Proc Natl Acad Sci U S A 91:12288-12292

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499-1504

Boeke JD, Pickeral OK (1999) Retroshuffling the genomic deck. Nature 398:108-9, 111

Boissinot S, Chevret P, Furano AV (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17:915-928

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331

Cost GJ, Boeke JD (1998) Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry 37:18081-18093

Cost GJ, Golding A, Schlissel MS, Boeke JD (2001) Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res 29:573-577

Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab 67:183-193

Deininger PL, Batzer MA, Hutchison CA, 3rd, Edgell MH (1992) Master genes in mammalian repetitive DNA amplification. Trends Genet 8:307-311

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH, Jr. (1991) Isolation of an active human transposable element. Science 254:1805-1808

Economou EP, Bergen AW, Warren AC, Antonarakis SE (1990) The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. Proc Natl Acad Sci U S A 87:2951-2954

Fanning TG, Singer MF (1987) LINE-1: a mammalian transposable element. Biochim Biophys Acta 910:203-212

Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905-916

Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci U S A 88:7396-7400

Gilbert N, Lutz-Prigge S, Moran J (2002) Genomic Deletions Created upon LINE-1 Retrotransposition. Cell 110:315-325

Goodier JL, Ostertag EM, Kazazian HH, Jr. (2000) Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum Mol Genet 9:653-657

Grimaldi G, Skowronski J, Singer MF (1984) Defining the beginning and end of KpnI family segments. Embo J 3:1753-1759

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361-372

Hammer MF (1994) A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. Mol Biol Evol 11:749-761

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979-988

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 94:1872-1877

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119-121

Kazazian H, Goodier J (2002) LINE Drive. Retrotransposition and Genome Instability. Cell 110:277-280

Kazazian HH, Jr. (1998) Mobile elements and disease. Curr Opin Genet Dev 8:343-350

Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332:164-166

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72:595-605

Miyamoto MM, Slightom JL, Goodman M (1987) Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. Science 238:369-373

Moore JK, Haber JE (1996) Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature 383:644-646

Moran JV (1999) Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay. Genetica 107:39-51

Moran JV, DeBerardinis RJ, Kazazian HH, Jr. (1999) Exon shuffling by L1 retrotransposition. Science 283:1530-1534

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr. (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87:917-927

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato T, Taccioli G, Batzer MA, Moran JV (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31: 159-165

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA (2002) A Comprehensive Analysis of Recently Integrated Human Ta L1 Elements. Am J Hum Genet 71:312-326

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, et al. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616-1622

Ostertag EM, Kazazian HH, Jr. (2001) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res 11:2059-2065

Ovchinnikov I, Rubin A, Swergold GD (2002) Tracing the LINEs of human evolution. Proc Natl Acad Sci U S A 99:10522-10527

Ovchinnikov I, Troxel AB, Swergold GD (2001) Genomic Characterization of Recent Human LINE-1 Insertions: Evidence Supporting Random Insertion. Genome Res 11:2050-2058

Perna NT, Batzer MA, Deininger PL, Stoneking M (1992) Alu insertion polymorphism: a new type of marker for human population studies. Hum Biol 64:641-648

Prak ET, Kazazian HH, Jr. (2000) Mobile elements and the human genome. Nat Rev Genet 1:134-144

Rothbarth K, Hunziker A, Stammer H, Werner D (2001) Promoter of the gene encoding the 16 kDa DNA-binding and apoptosis-inducing C1D protein. Biochim Biophys Acta 1518:271-275

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH, Jr. (1997) Many human L1 elements are capable of retrotransposition. Nat Genet 16:37-43

Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD (2000) Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res 10:1496-1508

Skowronski J, Fanning TG, Singer MF (1988) Unit-length line-1 transcripts in human teratocarcinoma cells. Mol Cell Biol 8:1385-1397

Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9:657-663

Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246:401-417

Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA (1997) Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res 7:1061-1071

Symer D, Connelly C, Szak S, Caputo E, Cost G, Parmigiani G, Boeke J (2002) Human l1 retrotransposition is associated with genetic instability in vivo. Cell 110:327-338

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD (2002) Molecular archeology of L1 insertions in the human genome. Genome Biol 3:research0052

Teng SC, Kim B, Gabriel A (1996) Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. Nature 383:641-644

Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R (1998) Apolipoprotein(a) gene enhancer resides within a LINE element. J Biol Chem 273:891-897

CHAPTER FOUR:

SUMMARY

Retrotransposons have had a major impact on human evolution by helping to shape the structure of the human genome and in the creation of genetic diversity. In this study the impact of L1 mobilization on the human genome was examined and the human genetic diversity associated with the recently integrated L1 elements was measured.

In chapter two, we analyzed 468 Ta L1Hs (L1 human specific) elements from the draft human genomic sequence and estimated that this subfamily contains over 520 members in the human genome. A total of 262 Ta L1 elements were screened with polymerase chain reaction (PCR)-based assays across diverse human populations and various non-human primate species to determine their phylogenetic origin and the level of human genomic variation associated with each element. One hundred fifteen (45%) of the Ta L1 elements were polymorphic with respect to insertion presence or absence. All of the Ta L1 elements analyzed by PCR were absent from the orthologous positions in non-human primate genomes, except for a single element (L1HS72) that was also present in the common (*Pan troglodytes*) and pygmy (*P. paniscus*) chimpanzee genomes. Sequence analysis revealed that this single exception is the product of a gene conversion event involving an older preexisting L1 element.

One hundred twenty-four of the elements were full length (6 kb) and have apparently escaped any 5' truncation. Forty-four of these full-length elements have two intact open reading frames and may be capable of retrotransposition. Sequence analysis of the Ta L1 elements showed a low level of nucleotide divergence with an estimated age of 1.99 million years, suggesting that expansion of the L1 Ta subfamily occurred after the divergence of humans and African apes.

The preTa subfamily of Long INterspersed Elements (LINEs) was analyzed in chapter three. Estimates suggest that this subfamily contains approximately 400 members in the

human genome, and has low level of nucleotide divergence with an estimated average age of 2.34 million years old suggesting that expansion of the L1 preTa subfamily also occurred just after the divergence of humans and African apes. Three hundred sixty-two preTa L1 elements were identified from the draft human genomic sequence, investigated the genomic characteristics of preTa L1 insertions, and also screened individual elements using PCR assays to determine the phylogenetic origin and levels of human genomic diversity associated with these L1 elements. All of the preTa L1 elements analyzed by PCR were absent from the orthologous positions in non-human primate genomes with 33 (14%) of the L1 elements being polymorphic with respect to insertion presence or absence in the human genome. Computational analysis of the preTa L1 elements revealed that 29% of the elements amenable to complete sequence analysis are essentially full-length (approximately 6 kb) and have apparently escaped truncation. Twenty-nine of these full-length elements have two intact open reading frames and may be capable of retrotransposition

Analysis of the L1 Ta and preTa subfamilies involved the characterization of over 800 recently integrated L1 elements from the draft sequence of the human genome. Estimates suggest that L1 mobilization alone is responsible for increasing the size of the human genome by roughly 1.4 million bases. The impact of these elements is even more dramatic when one considers the subsequent expansion of Alu and other retransposonable elements that were driven by L1 enzymatic machinery. Computational evidence suggests that over 70 human specific L1 elements may still possess the ability to retrotranspose within the human genome and include currently active L1 source genes that continue to fuel the expansion of retrotransposons within the human genome. Interestingly, over 35 Ta and preTa L1 elements were found adjacent to exons, though the majority of insertions showed a general preference for gene poor regions of the genome with low GC content. In addition, over 500 recently

integrated L1 insertions were analyzed using L1 element specific PCR assays. One hundred forty-eight of these insertions are polymorphic with respect to insertion presence or absence in the human genome, providing further evidence that L1 mobilization occurred during recent human evolution. Furthermore, the L1 insertion polymorphisms reported here will provide useful genetic markers to study human population genetics.

APPENDIX A:

LETTERS OF PERMISSION

The University of Chicago Press

Permissions Department
1427 East 60th Street
Chicago, IL 60637
Telephone: (773) 702-6096/Fax: (773) 702-9756

Gratis Permission Grant

Date: January 28, 2003

Grant number: 55834
Re: your request dated 1/10/03
your reference #

Dear Requester:

Thank you for your request for permission to use material from the publications(s) of the University of Chicago Press. Permission is granted for use as stated below. Unless specifically granted below, this permission does not allow the use of our material in any other edition or by any additional means of reproduction including (by way of example) motion pictures, sound tapes, electronic formats, and phonograph records; nor does this permission cover book clubs, translations, abridgment, or selections which may be made of the publication. No subsequent use may be made without additional approval. This permission is subject to the following terms:

1. On each copy of the selection full credit must be given to the book or journal, to the author (as well as to the series, editor, or translator, if any), and to the University of Chicago as publisher. In addition, the acknowledgment must include the identical copyright notice as it appears in our publication.
2. This permission does not apply to any part of the selection which is independently copyrighted or which bears a separate source notation. The responsibility for determining the source of the material rests with the prospective publisher of the quoted material.
3. This permission includes use in Braille, large type, or other editions of your work by non-profit organizations solely for use by the visually handicapped provided no fees are charged. The limitations listed in clause 4, below, do not apply to such use.
4. This permission covers publication of one edition of the work up to **250** copies.
5. Permission granted is non-exclusive and, unless otherwise stated, is valid **throughout the world in the English language only**.
6. Author approval is not required.
7. Permission is granted GRATIS.

To: Jeremy S. Myers
Louisiana State University
107 Life Sciences Building
Biochemistry and Molecular Biology Division
Baton Rouge, LA 70803

Material Requested

Myers JS, Vincent BJ, Udall H, Watkins WS, et al., "A Comprehensive Analysis of Recently Integrated Human L1 Ta Elements" AM JOUR OF HUM GENETICS 71: 312-26

To appear in the doctoral dissertation of Jeremy S. Myers, to be submitted to the Dept. of Biological Sciences at Louisiana State University.

Approved by: _____, **Perry Cartwright, Rights and Permissions**

Federal ID No. 36-2177139 250\1\$0.00

17 March 2003

Our ref: HW/mm/mar 03.029

Mr Jeremy S Myers
Louisiana State University

E-mail : jmyers9@lsu.edu

Dear Mr Myers

JOURNAL OF MOLECULAR BIOLOGY, *Vol 326, 2003, pp 1127-1146, Salem et al, "LINE-1 preTa…"*

As per your letter dated 27 February 2003, we hereby grant you permission to reprint the aforementioned material at no charge **in your thesis** subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

   "Reprinted from Publication title, Vol number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier".

3. Reproduction of this material is confined to the purpose for which permission is hereby given.

4. This permission is granted for non-exclusive world **English** rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form. Should you have a specific electronic project in mind please reapply for permission.

5. This includes permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

Yours sincerely

Helen Wilson
Rights Manager

Your future requests will be handled more quickly if you complete the online form at
www.elsevier.com/homepage/guestbook/?form=permis

APPENDIX B:

SUPPLEMENTARY DATA

Table 1 - L1Hs Ta PCR Primers, Chromosomal Locations, and PCR Product Sizes.

Name	Accession	Chr. Loc.[1]	5' Primer Sequence (5'-3')	3' Primer Sequence (5'-3')	A.T.[2]	Human Diversity[3]	Product Sizes[4]		
							Filled	Empty	Subfamily Specific
L1HS1	AC010739	2	AGGGAATGCTTATATTGTTGATGAG	ACTTCCTTCAGGGTTAATAGCAAAG	60	FP	3877	159	224+
L1HS2	AC010305	16	ACCAAATATCTGGACACTTTCTGG	GAAGTCAGCAGTGGTTAATTTTACA	60	IF	6131	74	171
L1HS3	AC008572	5	GCTTCTAGAATTGGAAGTAATATGG	AGTAGCCTTGAATCATCTTTTG	56	FP	656	95	422+
L1HS4	AC009494	Y	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	467	-	-
L1HS5	AC020647	12	TCAACTACAAAGTTGAAGAATAGG	GTTTCCATCAACAAGATCATGTCAAG	58	LF	546	376	455+
L1HS6	AC016138	3	TTTATTTCCCTGCATCTGATTA	CCTGTTATTAGATAATGAGTTCTAGTC	54	HF	402	122	219+
L1HS7	AC004773	7q11	CCTTAGACATATTCTTGGAAATAG	CCAGAATATTTGGGTATTTCATCTG	58	HF	326	169	256+
L1HS8#	AC004491	7q	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1689	-	-
L1HS9#	AC004694	7p	TCTTTCAATGGAAACAAGAGGTATC	AGGGAGAGGGACACTGAGTTTAT	59	FP	6126	74	178
L1HS10*	AF149774	7p	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6076	-	-
L1HS11	AL049842	6q	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	667	-	-
L1HS12*	AC007538	Xq28	GTTAAAGCAATCAAGCAATCTACTG	TAACAAGGCCACTGTAGAAAAGATT	59	FP	6188	104	209
L1HS13#	AC007938	7q31	ATGGGAAGGAACCCCATCTAT	AATTACTCCTCTCTTTGGCCTGTT	59	HF	745	128	220+
L1HS14	L05367	17q	AAGTGGATTAACAGTAACATACAGA	CCAAGCTGATAACTGATTATCTCA	55	IF	601	251	158
L1HS15#	AC007556	2	AATGCATACCCATGAGGACAA	ATGGTGTTGCACAACAAAAGAA	60	HF	6167	126	197
L1HS16	AP000220	21q	CCCTCACAGAGTGCTTGGTAA	GGGAAGGTAGGAAAACAGATT	56	IF	368	101	207+
L1HS17#	AC007486	X	GCATCCCTAAAGCAATAATCCA	GGAATTTTCCACTTGTGGTGTC	60	Paralog	4286	90	170+
L1HS18	AC005798	4	TTGAACAGCTTAGACTCGTCAGATA	GCAGTTAGACAGGAAAACAGAAAGA	60	HF	6174	87	212
L1HS19	AC007876	Y	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6115	-	-
L1HS20	AC009241	2	AATGGAAGAGCTCTCAAATTCCTTA	GCAACCATTCAAAAATTTACAACAG	61	IF	2302	62	181
L1HS21	AC008277	2	GTGTTGGCATATTTCTATTCG	TAAAGGCTGAACTTTGCATTG	57	LF	2606	84	178+
L1HS22	AC010682	Y	GCTCTCGGGTTCTTCTACCTCT	TCTACTGTTCCATGCAATAGATGTG	60	NR	3216	266	249+
L1HS24#*	AC004554	Xp22	GTGTATTTTGCCTTTTGAACCAA	CAAAAACTTGTTTCACTTGATTTTTAG	59	IF	6148	101	181+
L1HS25#*	AC002385	7q31	GAGGACCTTATTCATTTATTGC	CCATCTGAGCTTTAGTTTTGTCATA	60	FP	6140	94	191+
L1HS26#	AC003689	11q12	GCTTCAAGCTTAAAAGATGTAGACT	CCTACCCAAGTATCCACTGTCC	60	IF	2652	589	420+

(Table cont.)

L1HS27	AC007736	2	AGAACGTTGCCACATTATTTTGA	GTAGGAAGGTCTGGACTGGAGTATT	58	FP	3667	68	214+
L1HS28*@	AC002980	Xp22	CTTTTGTGACACTGGATTTCTAGC	CACTGTATATTGGAGCTGTTTTTCC	58	IF	6531	282	373
L1HS29#	AC005090	7p	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1476	-	-
L1HS30#	AL022166	Xp11	CCCTAAACAGAAAGGAAAATGAGAC	TCCTCATTGTGGTTCAAGGTTATAC	60	IF	4795	97	175+
L1HS31@	AC019212	X	GACAACACAAAGAAAACCCAAGAT	CTTATGTCCCAAAGCTAGTGAGTGA	56	FP	2317	86	176
L1HS32#	AC004911	7q	TCTCTAATCCAGCCTTTCAATTC	TGTTTCTTTTCCTGTGTGTTTCC	57	IF	463	280	384
L1HS34@	AC002122	5p15	ATGTCTGTCTTGACATTCCTAAGC	AATATGTAGAATGGCACAGGCTTC	58	IF	2177	284	328
L1HS35*	AC010081	Y	CTACCACATAACTGAGTGACAGTTT	CAATGTGCATCCATATAGCTGTGTT	61	FP	6308	233	239
L1HS36#	AC004000	Xq23	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6038	-	-
L1HS37	AC003080	7q31	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6017	-	-
L1HS38#	AC004142	7q31	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS39	AC005690	4	AGAACCAATCTTGCCCACAC	TGAGGAGTTTCTGAGTAACCTGGTA	60	HF	6337	155	189
L1HS41	AF222686	Xp11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1959	-	-
L1HS42	AC020925	5	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	580	-	-
L1HS43	AF172277	7q21	TTTATTGCACCTCCTGGTAAAGTAG	AGAGCACCATTAAACAACACAAGAT	58	IF	6157	89	191
L1HS44#	AC004883	7q	TAGCTGTGCTTGTTATGTCCAGTT	GAATGAGTTTTGTGTGGTTCTGTG	57	VLF	2288	478	615+
L1HS45#	AC004865	1	AATAGGCCCAGCTATTAGATTTAGC	CCTTTAAACCTTTGAACACGATTT	53	FP	329	81	150+
L1HS46#*	AC006027	7p	CCTGTGTTCCTTTTGTAATCC	CAAATGTCTCTTCAAGGACTG	55	HF	6382	326	183+
L1HS47	AC006986	Y	AGTCAAATGATTTTTAACTGCTG	GAGGGCAAGATCATGAAACA	58	Paralog	6177	86	230+
L1HS48#	AC005105	7p	CGAAAAGCTTAGGAAACTGTTTGT	TAAGCAATCTTCAGTTTAGGAAA	58	FP	1242	810	420
L1HS49	AC010202	12q	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	612	-	-
L1HS50	AF198097	Xp11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6308	-	-
L1HS51	AC008055	12q22	GCCCCTTACGTTAGAATAGAAAC	TGGATTGGTCCATACTACTGT	55	FP	1094	272	239+
L1HS55#*	AC004704	4q25	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6063	-	-
L1HS56#	AC005908	12p13	CCATTCATCAGCCATTTGCTA	GTGGCTTTAAAACAACGAGATG	59	FP	6545	459	494+
L1HS57#	AC006222	4	CAGCAAGACTCTGTCTCTAAAATGAT	GGACTTGAATTTGGTCTTGTTTCTA	59	LF	589	195	284+
L1HS58#	AC005939	17	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6101	-	-
L1HS59	AC003678	11q12	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2081	-	-
L1HS60#	AC006465	7p	GAAGTATGGAAATTGAGTCACA	CCCTAAGCTGTATCACTTTAAAACA	56	FP	445	104	246+
L1HS61#	AC002288	16p12	ACGTTTGTGCTTCACTCTAAGTTCT	CAAAATACCGGGATTATAGTTGTGA	57	FP	353	68	175+
L1HS62	AC006840	4	ATTAAAAGGAATGGACATGCAACAC	AATCTCAAAAGCTTCCTTGCACT	60	FP	6282	182	256+

(Table cont.)

L1HS63	AC023423	Y	AAGAAAGTGTTGTCAGAGAGTGTGA	AGGCCATTGGTCAGTCATAATTT	60	Paralog	6160	115	200
L1HS65#	AC004053	4q25	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1781	-	-
L1HS68#*	AC004200	6p21	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6242	-	-
L1HS69#@	AC004220	5	GGATGTTGATGATGGAGTCAGTC	TAACCATTTGAAACCATTAGAGGTC	60	FP	1410	76	180
L1HS70#@	AL049588	Xq	GTTCATTTGAGTGAGGGTACTGTCT	TAAGTCCCAAAAATTGCATCC	59	IF	3174	175	256+
L1HS72	AL133413	9q	CTGAGATGAGACAGCAGGTCTTC	TCTGCTGAGATTCTTCCATTTACC	60	FP	825	147	221
L1HS73	AC018822	3p	ATAAGGAGCCTAGGGAAGAACTTTT	CAAGCATGCCTGAAACATCTAT	55	HF	1126	462	162+
L1HS74*	AC011990	17	CTGGACGTATTTCTTACAGAGTTGA	CCCTAAGTTATTTTCCTTGAGGCTA	60	LF	6163	125	186+
L1HS76	U08211	X	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS77#@	AB020867	8p	TTCCTAAATGGCCTTACTATCCTTT	TCAGAAGTGCTAACAACTCTAGTAGGA	58	HF	990	78	233
L1HS78#	AP000084	21q22	TAGTACCTCCCTTAAAGAGCTG	GAGGAAAAGAAAAGTGCCTGATA	59	IF	374	107	175+
L1HS80#	AC017051.4	UL	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1823	-	-
L1HS81	AP000962	21q21	AAGTGTTATATATTGGAGCAATTC	ACAAGACAATGCCAATTTTAAGAGA	60	FP	848	148	401
L1HS83#	AJ001189	Xq12	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS85	AC008132	22q11	TTTGTATGCCTTGTGTTTTGTATTG	AGGAGAGTCTCATCTCCAGAGTTAC	58	LF	593	79	183+
L1HS86*	AL121825	22	GCAGTATCAGGAAATGCAATACAC	GGGATTCAGTCACCTTTATTAGACA	60	HF	6154	410	180+
L1HS87*	AL078622	22	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6065	-	-
L1HS91#	Z84572	13q12	ATACGTGCAAAACAGGAGATTTGA	TGTTTATGGTGAAGGATAAGTCTCA	59	FP	1619	78	167
L1HS92	AL022153	Xq	ACAATCCCTACTTCAGAAAGTT	CAACACTTTGATCATGAATAATAGCTC	57	FP	859	121	206
L1HS93	Z95325	Xq21	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	4882	-	-
L1HS94#*	AL031586	Xq	TCGTATGAATAACCTTGTGTTCTTG	TTTAGATCCTCGTCACTCAAAGTGT	57	FP	6250	151	264
L1HS95#	AL023284	6q	GGAAATTCTCAAGCTCAAGTTAAAA	CTTTTAAAGTGTGTTCTCACAGTGG	60	FP	717	119	320+
L1HS97#	AL030998	Xq	AACCAAACCCACAATCAGTAGAA	CTAGCTAAAGGTTTGCTATTTTT	58	FP	1640	182	407+
L1HS98#	AL022099	6p	ATCTGCATTGGGCCAAGTTTT	TCTCCTGTAAGACAGCACCATA	60	FP	1561	129	242+
L1HS99#	AL022726	6p	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6290	-	-
L1HS100	Z98754	Xq	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6161	-	-
L1HS101#	Z72519	X	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS102#	AL096677	20p	CCATTTGCCATAAATAAAGGCATC	ACTGTTACAAGTTTCCCCAAATGT	59	FP	6741	611	542
L1HS103*	AL121591	20	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6019	-	-
L1HS104#	AL096799	20	GAGATGTGGTTTTGTTTGAACTG	GCAGCTCACATAGTTTAGAGAAGAT	59	IF	6196	131	219+
L1HS106	AL117339	10	CTGACTGTTGAAACTTCTCCATTG	CAATAGACATGAAGGCATGGAAG	57	FP	3103	378	345

(Table cont.)

L1HS108*	AL031768	6p	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6091	-	-
L1HS109*	AL137191	14	GCCTTTCTATCTTTTGCTCTTGGT	GACACATACCAATTACAGGCAAAG	59	FP	6549	501	381+
L1HS110#*	AL078623	20	GGATTCTGACCTTATTCTAACAGCA	AGTTGACTGTTGGTGTTGATTGTGT	56	HF	6263	212	253
L1HS111#	AC002069	7q21	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	535	-	-
L1HS112	AC018755	19	AGGTTCCATCTCTAATACTGGATAA	TGATCACTTTGTTGTTAAGATGGAG	60	LF	1686	102	170
L1HS113@	AL133386	6p	AGTTTTGGCCTGAGAGAGAAGTAGA	GGTAGGCTAGAGATCCCTTCAATTA	55	FP	405	184	328+
L1HS115	AL132639	14	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	182	-	-
L1HS116	AC024610	18	CTGTGCACTTTTCCATATGTTTGAC	TCTAATCTATGGTGGATGCTCTTTC	56	FP	252	76	189
L1HS117#*	AC005885	12q	TGCAGTGTTCTATTTATGTCGTAGGT	CGAGAGAGGGAGGAAAGTGAG	57	IF	6629	535	176+
L1HS118	AC020599	4	ATGCCAGAAATACCTCTTTTACCTT	CTAAGTGCAATTCTCTCAGATTTTG	60	IF	6321	286	277
L1HS119#	AC005739	5	GGCTTATTTAGAGCACCTGGATTTA	GAGATCCAAAGCTTATGCTGTAAGT	60	FP	904	243	257+
L1HS123#	AC005350	5q	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	397	-	-
L1HS124#	AC004499	20q	TGACATAATTAATGGAGAAAACCAG	GAGATCCCTGTCCTTGTGTGAT	60	FP	749	515	373+
L1HS125	AF001905	Xq25	CCTCACGTTTCTCCACATTGTA	TTCTGGCCTTCATAGTGTTTTA	60	HF	332	96	169
L1HS126#	AC004784	19q13	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1552	-	-
L1HS127	AC004384	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	225	-	-
L1HS129#	AC003100	4q25	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1132	-	-
L1HS130	AL133320	1p	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6066	-	-
L1HS131	AL163152	14	TTGACTGTGTACTGCCAGTCTCT	GTAACCTACCAGTTTACAGTTACC	58	IF	381	179	212
L1HS132	AP001693	21	CCCTGATACACCAGTATATCTTA	GAAAAGAAAAGTGCCTGATA	56	IF	753	486	173+
L1HS133	AC008716	5	CATGGTGTCCCAGTGTTAAAAA	TATCTCTTACCTCTTCTTGCCCATA	59	FP	3351	821	738+
L1HS134	AF265340	16	CACAGTCAACTCAACCACTGAATAA	AAGGAGATGGAAGTAAGTGCAAAC	60	FP	751	433	603+
L1HS135	AL137804	11p	TTTTTGAAGGGAGTACAGTAATAGGT	GCCTTCCATAGTTCCTATTTGC	58	FP	6475	429	500+
L1HS136	AL157791	14	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	175	-	-
L1HS137	AL157879	5	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6057	-	-
L1HS150	AP000966	21q21	CAAGAACAACTGAAAAATGCAGAT	CCCCTCAGTCTCTGGTTACCTA	58	FP	642	89	141+
L1HS151	AC019205	6	CTTTGATCAGTTCTTGGAACTAGGA	CCTCTATGCCTTATTCATGCTTATC	60	FP	573	405	476+
L1HS153	Z84814	6p	CCAATTCACTTTGTCTCCTAGAAAT	AGTTCACGAAGTTGAAAGCTTATGT	60	IF	931	169	219
L1HS155	AC019050	2	TGGCATGTCAATATATACCTGAAGA	GGAAAACAGAAATAAAAGACGGACA	60	FP	7004	596	720
L1HS157#	ALO49842	6q	ATTCAAGTTCCAGTAAGCTGTGTTT	GAACTTTGGAAAATTCACAACTACC	60	HF	892	143	245
L1HS158#	AC008467	5	CAGCCCAGAGTAGTTCATGTTTT	GAAGGAAAAGGAGCTGCTTAGATA	59	IF	6194	147	207+

(Table cont.)

L1HS159	AC009976	Y	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1439	-	-
L1HS160	AL121938	6q	CTAAATAGGCAGAGGAAAGGAAAAC	TAAACTTCCAAGAGATCAGCACTTC	60	HF	1071	99	225+
L1HS162	AC009404	2	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	463	-	-
L1HS163	AL139114	9p	GGGACAGGGGTTAAGATTTTATTTT	AGTTCTCAACTGTAAAGGCAGTGTC	60	IF	2898	85	251
L1HS164	AB045357	1q	GGAAGGAAGTGGGGATAATAAGTAA	CCCAATTCAGTTTCTTCATTCTATG	60	FP	1507	193	267+
L1HS165	AC011666	1q21	CACAGTGATGGAGTTACAATCTTTG	GCTTTAAAGTCAGACAGGCTTGAGT	62	FP	1509	200	276+
L1HS166*	AC021017	8	TGCCTGAAATGCTATTGGTAGTATC	GTGCCCAGCCCATAATATAAA	60	IF	6204	102	251
L1HS167	AC018637	7	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2975	-	-
L1HS168	AC009492	2	CTTTTTCAAGGCCATCTGTGAG	AATCCTTACAATGAAAAGGGTGT	61	FP	666	97	180
L1HS169	AL118519	6q	TATTGAGGTGTAACCAGCATACAAT	CCACACGAAAGATATATGAATTGC	60	IF	6289	214	288
L1HS171	AL137145	10	GAAAGTTCATGAAAGTTGTGATGC	ACAAGAGAATCTATCTCCTGAAGAA	60	IF	6157	91	198
L1HS172	AL133479	9p	CTAAGATCAGTCACAGGCTTAATGA	CAGGTGCAAGTGGTTTAATTTTC	60	IF	1326	111	193+
L1HS173	AL359218	14	CACCATCTAGTGATTTTATGTTCTGC	AATAATCCCCATTGACTGTGTACTG	55	HF	319	123	217+
L1HS174	AJ271735	Xq	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	3252	-	-
L1HS175	AL136382	1p	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	717	-	-
L1HS176	AC025819	Y	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1522	-	-
L1HS177	AC017015	18	CAAGTTCCTCACCAAATGAAACTAC	TCCATTTTACTGATGTTGAATAGGC	58	HF	693	165	273+
L1HS178	AC023480	3p	GAATATTGAGCTTTCTTCACCTTT	CAAGCATGCCTGAAACATCTAT	60	HF	508	54	162+
L1HS179	AC017089	4	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	3573	-	-
L1HS180	AC009276	7	GGAGTGTAGAATACTGGGGAAAATC	CTTATTTCCCAATGAGCCCTGTA	56	IF	507	84	225+
L1HS181	AC025759	5	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1179	-	-
L1HS183#	AC000100	19	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS184	AL450108	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6094	-	-
L1HS185	AL157837	1q	CTGGCAGTTCCCTCAATGTAA	GAGTAGCTAGCAAAACAGGTAATGAA	60	FP	604	108	214+
L1HS186	AL359332	14	GGTCTAACAATATTCATGATGC	CCTCTTTTACCCTGTGAAGAAAAT	60	FP	6313	249	205+
L1HS187	AL357153	14	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6059	-	-
L1HS189	AL512407	6	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	907	-	-
L1HS190	AC073893	Y	TCTACTGTTCCATGCAATAGATGTG	GGGTTCTTCTACCTCTGCATAACT	57	NR	3243	190	331
L1HS191	AC007972	Y	TCCTCCAAGACCCTCTAAAATAAAT	TTTTGTCTTCCCTGAGTAAATTCTG	60	FP	2645	122	251
L1HS192*	AC018680	4	TTTCACTTTTTCTATGGTGATGAGG	CTTAGAATGTTACACTTTTCCGACA	60	FP	6218	155	196
L1HS193	AC018503	3	CTACAGTGGCATTTCTTTAGGACAA	TATACAACAGAACTGAATCACTGAC	60	FP	6296	239	288

(Table cont.)

L1HS195	AC044791	15	GCTTACATCTCAAATTCTGGTACCTT	TGTAAGAGCCAAAGCCTTTTAAACT	60	FP	1521	150	209
L1HS196	AC025263	12	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6071	-	-
L1HS197	AC027332	5	TGGAGTAGAATTCAAGCAAACTGAA	AGAGTTTATGATAGGTCCCCATTCT	60	HF	6226	97	260+
L1HS200	AC009892	19	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1686	-	-
L1HS202	AL391097	20	TTGTACCTATGATTTGTGTGATAGGC	GCTCTACATAAAAAGATGTTCACCA	60	FP	990	754	435
L1HS203	AL354750	10	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	152	-	-
L1HS204	AL157815	13q	ACTAGTTGATGACAAACTGGATGTG	GAGTGGCATAATCAATTGCTAGAGA	60	FP	647	126	182+
L1HS206	AL355382	6	GTTTGTCAAGTGACAGGAATCTCTT	GCTAAGTCATCAATAAGCCCCTAAT	60	FP	2704	154	186
L1HS207*	AL354861	9	CTTTGCATATCTCTGTCATCCTACA	GATGAGATCATTCACACACTTTCTG	60	FP	6208	164	170
L1HS208	AL354793	X	AACATTGGGAGAAGTTTGCAGTAT	CCAAGTTGTTAAGCACTCCATAGTT	60	FP	6639	570	689+
L1HS209	AL158159	9	GATGAGTTATCTTTGACGCTTTGAC	TGATAGATGAATGAGCTTTATGGTC	57	FP	508	118	213+
L1HS210	AL135908	6	ATGTGGGGAAGATGAAGAAATC	GAAAACCCCACTATAGGAGTAAATTG	59	NR	5322	132	564
L1HS211	AC079598	12	TCTATCGTCTCTGTCTTCTTAATGC	AATGACACTCTGCCTTCAGACTTAG	57	NR	3001	275	407+
L1HS212	AL157700	Xq	TTCTAGCCCTCTACTAATGTCCTTG	TTCTAAGGTAGCTGCAGATAAGTGG	60	FP	1045	184	234+
L1HS213	AC087432	3p	AATGCCTGATAAAAGTAGACACACC	GTGGGAATATATCTTCTTGGGTTT	60	HF	1710	89	188+
L1HS214	AC007483	3	TAGCTGAGAAACCATAAGCCTAGAA	ACCTGAATGTCCACTCATTCACT	60	HF	4159	328	330+
L1HS215	AC037423	9	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1162	-	-
L1HS216	AC023880	7	CTATACCAAATGCAGTCAGGATGTT	TCCCATAACTCTGTCACACTAGAAA	59	FP	714	197	228
L1HS217	AC073148	7	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6063	-	-
L1HS218	AC016910	2	TCTTACAGCACTATTCAGTGTTTGC	TTCCTCTCAAGGAACTCAAACC	60	FP	6136	82	174
L1HS219	AC021020	3	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6096	-	-
L1HS220*	AC016635	5	ATTGGCCTTCAGAAGTGATTAAGAC	TAGATAGCCAGACAAACAAACCTTG	60	LF	6244	135	260+
L1HS222	AL445932	6	TCTTTCTCCTCTTGTAATGTCTCAG	AAGATACTGTGCTTCACTCTTCTGG	60	LF	6195	118	238
L1HS223	AL450488	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	4210	-	-
L1HS224	AL358934	9	GATCTGAATCTTTGCTCTCCAGATA	ACGTGGTACAAAAGAAAACACTGTC	60	FP	1121	126	215
L1HS225	AL445523	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	3537	-	-
L1HS226@	AL353153	6	CCCTAAGCCTGTCAGAAGTTAGTATC	GCCATGAAAGATAAGGAGATAAGAG	60	LF	2114	120	359
L1HS227	AL157701	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	518	-	-
L1HS228	AL353657	13q	AATATCCACTACCCAATTCCATAGG	GCTGCAATTTAGCAGGATTTCT	60	HF	1383	184	205
L1HS230	AL359174	6	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1291	-	-
L1HS231	AL354896	13	GAGTATGAGAGCTCTGCTTTCTGTC	CTTGAAGGACTGGGATACTTGAAA	60	HF	2289	379	481

(Table cont.)

L1HS232	AL365367	1p32	TGTCACTCCAGTGATAGAAGCTAGA	ACAGTTAACTTCAAGGCAGGTTGAC	60	FP	1181	69	214+
L1HS233	AL357507	6	TAGTTGTCTACAACCAAGTGCTGAG	TCTGCATAGATCAGGAATTCTAAGG	59	IF	1232	81	174
L1HS234*	AL356438	6	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6092	-	-
L1HS235*	AL158193	13	ACAGGATCTTAAGGTTGAAGGTTTG	GGTTCTACCCAAAGTAGTCAAGAAA	59	IF	6441	420	179
L1HS236	AL365400	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1711	-	-
L1HS238	AL357519	6	GCAGGTAGGATACATGTAAGCATTT	ATCACAGCAATGGCATATCATC	60	FP	2155	374	360+
L1HS240*	AL137845	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6103	-	-
L1HS241	AP003112	8q23	GATAATCAGGTGATTGTGAACTGTG	CTACCACCCTTTTTACTCCCTTTAC	60	FP	366	148	206+
L1HS242#	Z80899	6p21	AGTTCACGGTCTCTATCTCTCCTTT	AACCTGTCTTTGACTGTTGAGC	58	IF	576	150	277+
L1HS243	AC019041	2	CACTAACATTCTGCATCTCACAATC	GTGGGAGGACATGAATAACACAT	58	FP	6148	96	202
L1HS244	AC009269	15	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	5512	-	-
L1HS245	AC017040	2	AAGGCTCTTTATCACAGGAAGTACC	ACGTTAATCACCGATCATTGC	60	FP	2141	294	263+
L1HS246*	AC068723	15q21	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6224	-	-
L1HS247	AC009274	7	GTGTGAAGTATTACCTCGGTGTTG	CTGTGTGGAGCAATAGTAACCAGAT	60	FP	2238	286	275
L1HS248*	AL360236	6	AGAACAAGTGAGTGGCTAAAACCTC	AGCCAACAATTTTCCCATCTC	60	FP	6705	658	710
L1HS249	AL355852	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1297	-	-
L1HS250	AL162373	13	AGTACCTGGTGAGTTCTCCTCAAC	GGTCTTTTGTGAGATGTCATACCTG	57	FP	2055	110	194+
L1HS251	AL445429	6	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	757	-	-
L1HS252*	AP002768	11q	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6026	-	-
L1HS253	AP001955	4q	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1780	-	-
L1HS254	AC013546	8	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	5961	-	-
L1HS255	AC022731	8	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1104	-	-
L1HS256	AC019218	8	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS257	AC016756	8	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS258	AC024905	3	GATTGGACTCCATTTCCTCTTGTAT	ATAAATTCTGGGACCTCTGCTTAAT	57	FP	1717	1011	643
L1HS259	AC020707	9	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1893	-	-
L1HS260	AL354982	9	GGCAACGGAATAATAGCTTCA	GTCAGCACTCCCATCTTAAATGTCT	57	HF	6461	358	510+
L1HS261	AL161631	9	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1904	-	-
L1HS262	AC013579	1	GATCCCTGTGTCTGGAGCACT	GGAATTCATGGAGAAGGTGAGTT	60	FP	1148	97	186
L1HS263	AL356139	9q	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	889	-	-
L1HS264	AL391643	9	GAGGAGGAAGAAGGCTGATAATATG	GACAGCCACTAAGTTAATGAGATCC	60	FP	284	133	174+

(Table cont.)

L1HS265*	AC018938	9	GCATTATTTCTGGAGCACTCACT	GTCTTGTGCTATTAAGCCTGGTCT	60	FP	6087	105	207
L1HS266	AL137021	9q31	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	207	-	-
L1HS268	AC025428	10	CTTTGCTCTCTTGCTCCATGTAT	TATCTGTTTACCAACCCATCTCACC	60	FP	6235	90	283+
L1HS269	AC020642	10	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS270	AC026989	14	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	313	-	-
L1HS271	AC020644	10	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS272	AL157787	10	CTATGTCCTAGCCTTCCCAGATG	AGAAAAGACAAGACAGGATAGGG	58	FP	1125	201	223+
L1HS273	AL354951	10	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS274	AC027118	10	GCACATGGCTTCTTAGCTAACTT	CTTTCTTGCATAAATGACTCTGTCC	57	FP	2081	611	317
L1HS275	AL590378	10	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1414	-	-
L1HS277	AC026393	10	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	312	-	-
L1HS278*	AC027591	11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6020	-	-
L1HS280	AC078971	11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6063	-	-
L1HS281	AC037434	11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	343	-	-
L1HS282	AP001002	11q	CTTACCTCCAGAGCATGCACATTAT	CCCCTCCTTCTCAATTTAAGGTTAC	61	FP	6448	156	249+
L1HS283	AP000409	11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2294	-	-
L1HS284	AC018619	11	AGATAGGAGAATCCTCTGGTCTTCT	CTATTGTTGGGTACTTGGGTCACT	58	FP	1877	174	268+
L1HS285	AC015772	11	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS286	AC011829	11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1189	-	-
L1HS287	AC021304	11	CCTTTTATCTGAAATAAGTGGTTGG	CTTCCTTTAGCTGGGCTGTTCTAAG	61	VLF	1693	95	216+
L1HS288*	AC016775	11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6081	-	-
L1HS289	AC021245	11	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS290	AP001179	11q	CCTGTCAGTCTTATCTTTGCTCTACA	GGCATAGAGACAAATCCAAATTAAG	60	NR	6537	285	235
L1HS291*	AC025410	6	CTCCCACTACTTTATGGGAAGGT	AGGACTTCCAATTCCTAGTATGCAG	58	HF	5658	216	271+
L1HS292	AC073915	12q	GACTCCACACTAGCTTCTTTGACTT	GAGACTCAGTTGACAAGGAGTTACC	60	FP	1117	117	213
L1HS293	AC026831	12	TTACAATGGATACGTTAGACAGCTC	CCATAATTGGTTAGGATGATGAGAC	60	LF	2517	417	317+
L1HS294	AC027442	12	CTTTACCTGTTCCACTAATCAC	GGCACAAGATGGATATAAAGGA	57	FP	6154	103	168
L1HS295	AC012144	13	GAGGAATGGTTGAACAGCTTG	ATGTGGCTGGAGAAATACCTCTAAG	61	FP	713	100	208+
L1HS297@	AC064857	12	GTCCAGAGTGATGCATTTTATTTGG	GCATAGTCATTTAATGCATGTCAGC	58	FP	771	461	549+
L1HS298	AC025880	12	ATATACCATACTCCTTTCCCCTTCC	TGAGCCCTGTATTTTAATCACTTGT	60	LF	1037	80	235+
L1HS299	AC027287	12	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-

(Table cont.)

L1HS300	AC026577	1	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	3364	-	-
L1HS301	AC027382	1	CTATCCCATAGATGGTGGGTAGAAT	GAGGAAATAGCACAGGTATGGTAAA	61	IF	1770	411	431
L1HS302	AL365220	1p21	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2391	-	-
L1HS303	AL451063	1	CTATGTTCTGGGAGAAGAGCTGAT	CTAGGGTCAGAAAGAACTTTGATGT	62	FP	780	87	170
L1HS304	AL354885	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS305	AC016371	1	CAAAAAGCAGCCCTATATTAGC	GCCTGCCTCATTATCTTTCATT	58	FP	3998	415	409+
L1HS306	AL136459	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS307	AL390860	1	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6066	-	-
L1HS308	AL390200	1	CCTACTAGGCCCTCTTCTTTTGTAT	GTCTTGTTGTGCCAGACACTTTA	62	IF	3441	455	652+
L1HS309	AL391904	1	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2161	-	-
L1HS310	AL157946	1p31	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	286	-	-
L1HS311	AL162402	1p13	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	693	-	-
L1HS312	AL139225	1p13	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	783	-	-
L1HS313	AC034157	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS314	AL357975	1	TGGCTAGCAAAAAGGTGGAC	AGGGCAGAGAAAAATGGTCA	58	IF	6215	109	255+
L1HS315	AL139137	1	AAGTCCCAATTCCCTAGTCTGTCT	GACACAGAATCATGTCACAATACCC	61	FP	6286	77	332
L1HS316	AC026905	1	CTTTAGCAGTTTTCATGCCTCCT	AGGTTGATGGTAACCTGTAGGAAC	59	FP	6240	173	245
L1HS317	AL356323	1	CTCTGCCTCAAGTGTGTCTTGACTA	GAGAACACACCCTTGCTCAGTAAAT	59	FP	901	711	626+
L1HS318	AL365225	1	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	5243	-	-
L1HS320	AL357973	1	GGGATTCAAATGGGAAACAAG	CTCCTTTCCAGTATCTGCTCTTATG	60	IF	1748	140	305
L1HS321	AL356455	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS323	AC068071	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS324	AL139284	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS325	AL360154	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS326*	AC025702	1	CTCACCGTTATCAAAGGGTAGAAAC	CTAGCCCCAAATTTGAGAAACAG	60	FP	6250	156	289+
L1HS327	AC018874	1	GGTACAATGTAATCATGGGTTGG	GAGTTAACCGTTAGTCCACAAGATG	58	FP	4695	172	413
L1HS328	AL135842	1q21	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2188	-	-
L1HS329	AC058795	1	CTTCACCTCTGAATGACACACAT	GGCTTCATAATGCATCGCTAA	60	FP	1188	454	365+
L1HS330	AL139285	1p31	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS331	AL138777	1q31	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1064	-	-
L1HS332	AC008110	1	CATGTTAGAACTGGCTCAAGTATCC	CCTGCAGAAATTTGCCTTTAG	58	IF	2850	87	227+

(Table cont.)

L1HS333	AC023026	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS334	AC026253	2	ACACTTCTGAGAATTTCCCTGTG	TTACTCCCTCTTTACTGTCTTGGTG	60	FP	1095	199	341
L1HS335	AC023434	1	CATGCATCTCTGAACTACTGACTTG	ATAAAAACCTGTTTAGGCCAAGG	60	IF	1276	395	284+
L1HS336	AC013264	1	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS337	AC010890	2	GGTACAATATGAGGCATCACGTA	GTAGCATCCTTTATAGCTTTGCTGA	60	HF	3174	210	329+
L1HS338	AC068953	2	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS339	AC017035	2	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS341	AC069384	2	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS342	AC018591	2	GAGACTCAGTTGACAAGGAGTTACC	AAACAGGACCTGCTGTCCATAA	60	FP	1087	78	183+
L1HS343	AC068572	2	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS344	AC048375	2	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS345	AC073509	2	CACAGCATTTACCAAAGCACTC	CTCAGTTCATTGCACAGTTTGG	60	LF	2587	192	229+
L1HS346	AC016674	2	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS348	AC018378.3	2	GAAATGGGAAGAGGAGTTGACA	CCTATTTTTATCTCAGCTGATGTCG	60	HF	748	283	526+
L1HS349	AC009963	2	GGAGCTGGGAGAATTATTGAAAC	CCACTCTCAACTACTGTCCAACAAG	60	HF	229	114	182
L1HS350	AC022605	2	TGGTATATAGTTCTAAGGACCCACAG	GCTACTTTTGCTTCTGGGTGTT	58	FP	725	243	331+
L1HS351	AC013262	2	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS352	AC073874	2	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	970	-	-
L1HS353	AC019324	2	TCCATGATAGAACACACTCTTCC	AATCCCTGTCAAAACCAATCC	59	HF	1822	426	167
L1HS354	AC012442	2	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6217	-	-
L1HS355	AC011901	2	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6067	-	-
L1HS356	AC009290	2	CATCCTGTTGAAGAACAGAGAGATG	ATAGAGTGACCAGAAACTCCAGAGA	60	FP	6290	156	250+
L1HS358	AC019130	2	GAGACTCTTTGGACTCAGAGTATAACC	AGTCCTGTCATACCAGTTATTGGAC	59	FP	6621	128	673
L1HS359	AC024062	2	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	4808	-	-
L1HS360	AC023416	2	GAGGTCTTTGTGCAGAGGTATAAGA	CTCACCAACATCAGTTTCCTTTG	60	IF	3222	153	218+
L1HS361	AC073642	2	AGCCCATTAGATATATGTGGCTGT	CTTTTTATATTGGTCACCCCCAAC	61	FP	6319	281	372+
L1HS363@	AC010913	2	GTTAGACAGCGACATGCACAG	ACCTCTGTGCCTTACCAAAAAC	60	FP	577	106	198+
L1HS364	AC026860	3	CTTAGCCTCTGTCTTTAGGGAAAAC	CATGACCAACGGTGCATAATA	60	HF	6139	97	170+
L1HS365	AC068355	3	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	888	-	-
L1HS366	AC083853	3	AGAAAACTTCCAGACACCTATCC	CTATGTCCTAGCCTTCCCAGATG	60	FP	1088	163	183
L1HS367	AC078805	3	GACTCATATTACCCTGGACAACAAC	AGTCTCTCCTTGCTCAGTTTGGTAG	60	FP	6784	83	401+

(Table cont.)

L1HS368	AC023144	3	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	168	-	-
L1HS369	AC076971	3q	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS370	AC068365	3	GCAATCAGTTTCACACTCAACTG	CATGTGATCTATTGTGTACCATCAGG	58	FP	3436	146	323+
L1HS371	AC026611	3	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS372	AC022077.13	3	GAAGAGAAAGAGGAAATAGCACAGG	CTATCCCATAGATGGTGGGTAGAAT	60	IF	1779	599	431+
L1HS373*	AC022838	3	GAAAGAGAGTTCTCTGTACCACACC	GTCATGTCCCAACAGGACATTT	60	VLF	6294	215	231
L1HS374	AC063919	3	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6265	-	-
L1HS375	AC023139	3	TGTGGTACAGTCACACTACAAAG	GATAGCATACACCATCATGCACT	60	IF	3862	430	469+
L1HS376	AC069203	3	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS377	AC078856	3q	GGGAGATGTAGAGTTTTATGTGACC	CTAATGTGCTGGGCAAACATAAGAT	57	FP	577	139	201
L1HS378	AC069225	3	CTCCCCTTTTTGCCTTACTTCT	CTTACTTGCAATAGCCCATTCAC	60	IF	5569	646	369+
L1HS380	AC024470	3	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS382	AC055732	3	GCAGACACTAGAAGCTTTTGCAT	GCCACAAAATCTGGCACTTATAG	58	FP	3357	426	185
L1HS383*	AC017085	3	ATTAGTCAGTAATAGAGCCCCCTGT	AAAGACTTCTTTCCAGCTCTACCC	60	FP	6493	267	515
L1HS385	AC078808	3	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6068	-	-
L1HS386	AC023438	UL	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS387	AC069417	3	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS388	AC025818	3	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	713	-	-
L1HS389	AC024216	3	CATGTAGAGATGATCTTCAAAGCTG	GCCTGATAAAAGTAGACACACCTG	60	FP	1782	162	263
L1HS390	AC036128	4	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS391	AC022040	4	GTGGACATCAGAGTATCCCTTTCT	AGAAGGGTACATGACAACTGGTTAG	60	HF	889	113	203
L1HS393	AC013336	4	TACACAGAATCTGATGCTAGGAGAG	CGGGAACATAAAGTCATAGCGTAAC	61	LF	751	277	412+
L1HS395	AC067804	4	GTTGCATTTTGGAAAGGAAGG	TAGTGGAAAGACAGACAGTTTAGGG	61	IF	1218	119	214
L1HS396	AC007512	4	AGACTCAAACTCAAAACTCCTGTGT	TCACAAGCAGACATTTCTTACTGAA	60	FP	6643	562	373+
L1HS397	AL161439	6	ACTCATCCTAGAGCTTTACCCAGTT	CACAAAGTCAACAGGTTTGATCC	58	FP	1085	259	231+
L1HS398	AC069349	8	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS399	AC027502	4	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	614	-	-
L1HS401	AC068037	4	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1342	-	-
L1HS402	AC020593	4	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	361	-	-
L1HS403@	AL158816	6	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	360	-	-
L1HS404	AC021700	4	CCACCTTACGTTCAGCTGTTAAT	CGGTGATTAGGTGACAGCTTTT	60	LF	3262	163	231+

(Table cont.)

L1HS405	AC032017	4	ATCAAAAGTCCTGTGTGTTTGTCTT	GAAATTTTGCTAGACATAGCTGTCC	60	FP	1206	396	202+
L1HS406	AC067842	4	GCAAGTTTTACCCATAGTACACAGG	GTATGTAGAAGGCAGGGGTACACT	60	HF	3589	209	302
L1HS407	AC041010	4	CTCACCAGTACGAGAAGCAAGTT	TCTGACCTAGGGATGATTCTTCA	60	FP	413	227	217
L1HS408	AC019133	4	TTTTAGCCAAGCTCTTTGTTCC	CATTATGGCAGCGTAGACATTG	56	FP	2059	106	209
L1HS409	AC027782	4	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS410*	AC011633	4	GCTAAGCAATGGAGGAAAATATCG	TGTACATGGTGTGAGGTATGAA	57	IF	6211	100	244+
L1HS411	AC073338	4	ACACACACACGATGGAAAGTATCT	AGCACATCCTAAATCTTCCTCTCT	60	FP	2670	136	246
L1HS412	AC067901	4	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS413*	AC023332	4	TCATGAGCATCACTCTTACCATGT	ACTCAGCTGACTTGCCATAAATGT	60	IF	6199	127	191
L1HS414	AC025955	4	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS415	AC009816	4	TCAGACCCATATATGAGCATAACC	GCTTAGAAGAATTTTTAGCCAGGTG	56	HF	1360	590	476+
L1HS416	AC068256	4	TTAGTCACTATGACTTGAGCCACTT	TAGTGATAGTGTAGAGAGGGGGTTG	61	FP	822	238	284
L1HS417	AP001860	4	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	865	-	-
L1HS418*	AC011981	2	CGATTTCTGTCTTTGTGAACGTAGT	CCTTACAGAGTAGAAATCTCACGAT	60	IF	6380	328	358
L1HS419	AC061978	4	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6034	-	-
L1HS420	AC041038	4	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6066	-	-
L1HS421	AC024974	UL	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS422	AC009577	4	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS423	AC022672	11	CTCCCTGTCTTCTGGGTTAAAATA	GGAAGTCCCACTTTTTCAGTAGAG	60	HF	5680	201	248+
L1HS424	AC080124	4	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS425	AC013724	4	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6120	-	-
L1HS426	AC023921	5	AGATTCCCTTTGGTATCCAAATCAC	GTTGCCATACTCCGCATAAAGTC	60	IF	3394	204	252
L1HS427	AC015990	4	TACGGGCAAAGACTGAGAGTACTAA	TTCAGCCTTCTGACATCAAACT	57	IF	2230	139	220+
L1HS429	AC060816	4	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS430	AC024963	4	CAGAGAACCAACATGTAGGAACAA	GTTACAGGTCAAAGGAGGTCTGAG	60	LF	4034	127	223+
L1HS432	AC011399	5	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS433	AC027339	5	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS434	AC010437	5	ACCTGGGCCACATTTATTTTTC	TGTAGAAGAAGACACCGTCGTTAG	60	FP	2637	250	246+
L1HS435	AC026403	5	GACTCAGTTGACAAGGAGTTACCA	ACACTAGCTTCTTTGACTTCACCA	55	FP	1115	111	211+
L1HS437	AC023526	5	ATCTATCATTTATCTGCCCCGTCT	ACAAGGATTAGCAGGAAGTCTGTT	60	IF	2954	256	201+
L1HS438	AC011433	5	TCCTCTCACCAACCACATAAAGTA	ATCCCTTGGATACAAAGATGTGC	60	FP	1909	570	345

(Table cont.)

L1HS439	AC016573	5	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS440	AC010409	5	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6133	-	-
L1HS441	AC026444	5	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS442	AC027325	5	GACGGTTACTCAGAAAAACACAAG	GTAGATGCCACTGTTACCCTGACT	60	IF	907	224	185+
L1HS443	AC021600	5	GCTAGACTCTCTACCTTTGGCTTT	TGATACCTGACTCTATGCACCACT	56	FP	891	261	382
L1HS444	AC027315	5	TTATTGGAATAGCTTCTCCTGTCAC	GCTGTTCCTAACTCTAGTCCTCCA	60	FP	464	303	296+
L1HS445	AC008374	5	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	551	-	-
L1HS446	AC010314	5	CTCGTGACATTTCCATCATATAGC	TTAAGTCACCTAAGGGTTGTAAGTG	56	LF	6142	109	182+
L1HS447	AC018759	5	GTACATCTCTTTGGACACTTCCACT	GTTTAAGTCCAACATCCTGTTCTG	59	IF	691	560	386
L1HS448	AC016545	5	GTCAATTAGAGCATGAAGAAACCAC	GTACATCTCTTTGGACACTTCCACT	60	IF	652	525	382+
L1HS449	AC011378	5	CTAGGGAGGTGAAAATTCAGATGT	GCATGTTGCACAACAGTATGTA	60	FP	1797	281	315+
L1HS450	AC011413	5	GTGAAGACTGTTGGTCAGTTACTTGT	GTCATTGAGATTGGCAGGTAAAAG	60	HF	6179	128	189+
L1HS451	AC010490	5	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	994	-	-
L1HS453	AL360232	6	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6064	-	-
L1HS455	AC027643	6	CATACACAAGGGCGAAGAGTTAAA	GCCTCTTTTACATCAGTTACCACTC	60	FP	259	110	213+
L1HS456	AC026966	6	TAACACTTAGTGATTGCTGGGAGAG	GGACAAGGTGAAGTGGAAAACTAGA	60	FP	1641	121	215
L1HS457	AC025887	18	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	286	-	-
L1HS460	AL355489	6	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6044	-	-
L1HS461	AL358992	6	ATCCAGCAAAAGTATCCCTTAAGTA	TCCTGTCCCAATTCTTTGTATTAT	60	LF	4143	324	417
L1HS462	AC069403	11	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	4163	-	-
L1HS463	AL391336	6	ATTAAATCTGTGTGGGAGTGG	AGGGTGACTTCAGTGATATCTTCA	60	FP	6304	247	346
L1HS465	AL356601	6	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1936	-	-
L1HS469@	AC020586	UL	GGTACTGGCTGTTCAGTATTTTT	GTCTCAAAGCCCATTTCATAGTTC	60	FP	6458	101	212+
L1HS472	AC018400	UL	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS476	AC079756	7	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	897	-	-
L1HS477	AC024730	7	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1271	-	-
L1HS478	AC069008	7	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	991	-	-
L1HS479	AC079855	7	CACTCGAAGGGTAAGTGAGATTTT	CCACTAGCGCACCATTTTTCTAAT	58	FP	6223	146	276
L1HS480	AC021836	4	AGAGGTAACCACTACCTTGCAACT	GCCTCATGACAGGAGAAGAGATAAA	60	IF	2701	272	265
L1HS483	AC026011	8	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS484*	AC073647	7	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6692	-	-

(Table cont.)

L1HS485	AC027189	8	CTCAGTTCCACATAAACCTTGACA	GAAGCAATTAACCTAGCAGTAGGAC	60	FP	548	74	183+
L1HS486	AL356516	9	CCCTCATCACCAAATATCTGAGAA	AGCTGACAGTCTAGTGAATGAGGTC	60	IF	905	139	196
L1HS487	AL162731	9	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6079	-	-
L1HS488*	AL353649	9	CAAATTGTCAATGCTAACCACTCC	GGAAAAAGGCACTTTGGCTTATC	62	FP	6787	724	472+
L1HS489	AC009284.2	9	TCTCCAGAAACCATCACAGTAAGA	AGGAGTTGAAAGTAGGATGGGTTT	60	FP	322	104	202+
L1HS490@	AL358937	9	CAGCTGTCTTGCTAAGAATCCAT	AGACCACAGACTCTTTGAGGGTAAG	60	FP	2289	397	206
L1HS491	AL355303	10	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS492	AL450466	10	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS493	AL138764	10	GACTACCTTTCTGCGTATTCCTTTC	GTCTAACAGGTACACGAGACTCCAT	61	IF	1603	111	241+
L1HS494	AC068972	8	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2974	-	-
L1HS495	AC083848	8	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1341	-	-
L1HS496	AC024929	8	CCTTTGGAAGAGAAAGAGGATATG	CTCCCAATGGAAAGGAACTTGTAT	60	FP	617	70	177
L1HS497	AC060775	8	GCCTAGTGGGAAGACAAAAAGTATT	GCTGTAATGTTAACCTCGAAGTCGT	60	FP	950	346	439+
L1HS498*	AC067844.3	8	AGGTTTCCCCAAAATTTACCC	CTGATGTGTGGATTCACTGTTCTT	58	FP	6281	184	295
L1HS499	AC024649	8	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1045	-	-
L1HS500	AC009630.5	8	GTGTTGCCTTCACCACAATAGTA	TTTCTCCGAGTACAGGTTACGAG	60	FP	1145	206	227+
L1HS501	AC022207	12	GTTGGCAACTTACTCTCAAATGG	AAATACACTCGACTGGCCACTAA	60	FP	6254	199	306+
L1HS502	AC011881	UL	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	537	-	-
L1HS503	AC055118	13	GTGAGGAATGGTTGAACAGCTT	TGTGGCTGGAGAAATACCTCTAA	60	FP	713	101	206+
L1HS504	AL158045	13	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS505	AL162716	13	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	384	-	-
L1HS506	AL138684	13	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS507	AC064832	15	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS508	AC048381	15	ACAGAACCTTTTAGAGGGAATCG	CTCCGTGTGGTAAAATTAGCTGT	58	HF	6144	103	184
L1HS509	AL356017	14	CACTCATGACTGCCTGACTTCT	CAGGGATTACTCTTCTGTTGTGG	61	FP	443	131	220+
L1HS510	AL390800	14	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1837	-	-
L1HS511	AL162632	14	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6088	-	-
L1HS512	AC021839	14	AAAGAGACAATCCACAGCATAGTTG	GATTTATTCCTTCATGGAGATGTGC	61	HF	2071	722	266+
L1HS513	AL160156	13	CCAAACTTGAGCCTCCTGTAATC	CCTTGAAATAAGCAGGAAGAAGC	61	IF	809	142	235+
L1HS514	AL138961	13	CCTCAGCTTTGGATCCTGTAGTT	AGAAGAATTGGGTCCTGTTGAA	60	FP	6670	334	361
L1HS515	AL163537	13	GGATGGTAAAGGAGTGGCATAAT	TGTGGAGCCCAGATCTTTTAAT	60	FP	637	106	193

(Table cont.)

L1HS516	AC044907	15	CCACAGTTTACACAGAAGCTGAA	GAAGGAGTGGATGTGTTTCAGTAA	60	IF	6151	101	212
L1HS518	AC074236	15	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2636	-	-
L1HS519	AC074100	15	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS520*	AC015558	15	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6087	-	-
L1HS521@	AC067951	15	GCTTTGTTTACCTTTCTGCTCACT	CACCAAAAGGAGAAGCCAATAAAG	60	FP	1248	344	441+
L1HS522	AC009555	15	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	190	-	-
L1HS523	AC009658.6	15	CGTGGAAGATGTTACGAGGATTA	AGAGAATGCGATGTCGATTAGAG	60	FP	570	105	204
L1HS524	AC020892	15	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS525	AC009057	16	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS526*	AC025289	16	ACCCTCCAAGGTAACTGAATCTTA	ATGCCCATGCTTGTTAGCTACTAC	60	IF	6076	223	324+
L1HS527	AC026472	16	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1224	-	-
L1HS528	AC009021.4	16	CGGATGGGAGCACAAAATTACTA	TGCCTACTAAGATACCTTGGAAATG	61	FP	991	172	278
L1HS529	AC022164	16	TGAGTAATGTGGCGGTTTAGTTC	AACCAGTCAAGAAGCCAAAGAG	61	FP	6143	116	193+
L1HS530	AC009063	16	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS531	AC055852	17	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	2839	-	-
L1HS532	AL356138	20	CCTCTAATCTATGGTGGATGCTCT	TGGTAGGGAGCTGGTAAAAGTCTA	61	FP	308	175	242+
L1HS534	AC007448	17	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS535	AC034266	17	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS539	AC034266	17	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS541	AC068204	18	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS542	AC023983	18	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS543	AC009267	18	TACATTAGTCTGCCTCTGATTCCA	GGCCATTCTTTTCATCTGTTGTAG	61	FP	547	99	183
L1HS545	AC007768	18	TGGGAACTCATGTTACAGTTTCAC	ATTTGTCATGATCACAGCCACCT	59	FP	2514	95	216
L1HS546	AP001460	18	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS547	AC010966	18	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS548	AP001113	18	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6237	-	-
L1HS551	AC021325	18	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	184	-	-
L1HS552	AP001564	18	CAGTGAACTGCTTTCTCACAATTC	CAAGAAGTTTTCCTGGAGTCTCTC	60	IF	4144	123	235
L1HS554	AC027230	18	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	561	-	-
L1HS556	AC026898	18	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS557	AP001019	18	ACAAAAGCACCTAGAAGCAGTCAT	CTTTTTCTCCTATGCTCGTGGTAT	60	FP	2277	85	229+

(Table cont.)

L1HS558	AC015819	18	TGCTTTCTTTCTTTCACATAGATCA	GCAGACACGAATCACAGTTTGTAT	61	HF	983	128	203+
L1HS559	AC023394	18	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1620	-	-
L1HS561	AC013620	14	TACCCATTTAAAGGGCAAAGTG	CTACCCATTTAAACCACTAATGCTG	61	LF	430	114	239+
L1HS562*	AC019175	X	TGTCTGTTCAGTCCTTTCTCACAT	AGCAAAATGTATGCCGAAGACT	59	FP	6170	115	181
L1HS564	AC034155.5	X	TGCAATTGACATAGATACTGCAGAG	CCCTTCCCTTTCTGTACATGTCTT	61	LF	2085	471	425+
L1HS565	AL442646	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	6029	-	-
L1HS567	AL158143	X	END OF SEQUENCING CONTIG	END OF SEQUENCING CONTIG	-	EC	-	-	-
L1HS568	AL356003	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	1297	-	-
L1HS569	AC021992	X	INSERTED IN REPEATS	INSERTED IN REPEATS	-	R	596	-	-

***** Full length elements with intact open reading frames.
# Elements previously identified by Boissinot et.al. 2000
@ Elements previously identified by Sheen et.al. 2000 and Ovchinnikov et.al. 2001

1. Chromosomal location was determined from Accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples.
2. Amplification of each locus required 2:30 min @ 94°C initial denaturing, and 32 cycles for 1 min 94°C, 1 min Annealing Temperature (A.T.), and 1 min elongation at 72°C. A final extension time of 10 min at 72°C was also used.
3. Elements at the end of sequencing contigs are denoted (EC), those residing in other repeats (R), those having paralogs (Paralogs), and elements with inconclusive PCR results (NR). Elements represented here are classified according to allele frequency as: high frequency (HF), intermediate (IF), low (LF), very low (VLF) or "private" insertion polymorphisms, or as fixed present (FP) insertions. Fixed present: every individual tested had the LINE element in both chromosomes. Low frequency insertion polymorphism: the element is present in no more than 1/3 (33%) alleles tested. Intermediate frequency insertion polymorphism: the element present in more than 1/3 (33%) of alleles tested and no more than 2/3 (67%) of the alleles. High frequency insertion polymorphism: the element is present in more than 2/3 or 67% but not all alleles tested. Indeterminable data is denoted (-).
4. PCR Product Sizes: Empty product size is calculated computationally by removing the L1Hs Ta elements and 1 direct repeat from identified filled site. Subfamily Specific product size is calculated from internal subfamily specific primer located in the 3' UTR to the proximal 3' primer. In cases where target site duplication sequence were not found flanking the element PCR product sizes may vary from those reported. Elements with subfamily product size denoted "+" were found in 5′→3′ orientation in GenBank and are

assayed using the internal subfamily specific primer and flanking reverse primer. All other elements were assayed using the internal subfamily specific primer and flanking forward primer.

Table 2A - L1Hs Ta autosomal associated human genomic diversity.

	African American					Asian/Alaskan Native[d]					German Caucasian					Egyptian					
	Genotypes					Genotypes					Genotypes					Genotypes					
Elements	+/+	+/-	-/-	**f[c]**	**H[a]**	+/+	+/-	-/-	**f[c]**	**H[a]**	+/+	+/-	-/-	**f[c]**	**H[a]**	+/+	+/-	-/-	**f[c]**	**H[a]**	**AH[b]**
L1HS2	1	7	11	0.24	0.37	11	6	0	0.82	0.30	8	9	3	0.63	0.48	7	7	2	0.66	0.47	0.40
L1HS5	0	2	18	0.05	0.10	0	2	18	0.05	0.10	1	7	12	0.23	0.36	0	6	12	0.17	0.29	0.21
L1HS6	17	1	0	0.97	0.06	18	0	0	1.00	0.00	18	0	1	0.95	0.10	14	0	0	1.00	0.00	0.04
L1HS7	17	3	0	0.93	0.14	19	0	0	1.00	0.00	19	1	0	0.98	0.05	19	0	0	1.00	0.00	0.05
L1HS13	15	0	0	1.00	0.00	15	0	0	1.00	0.00	18	0	0	1.00	0.00	18	1	0	0.97	0.05	0.01
L1HS14	9	11	0	0.72	0.41	7	9	3	0.61	0.49	1	11	8	0.33	0.45	2	9	9	0.33	0.45	0.45
L1HS15	13	4	2	0.79	0.34	20	0	0	1.00	0.00	18	2	0	0.95	0.10	15	5	0	0.88	0.22	0.17
L1HS16	1	6	13	0.20	0.33	7	9	3	0.61	0.49	3	6	11	0.30	0.43	1	3	11	0.17	0.29	0.38
L1HS18	19	1	0	0.98	0.05	19	0	0	1.00	0.00	20	0	0	1.00	0.00	18	0	0	1.00	0.00	0.01
L1HS20	3	15	2	0.53	0.51	9	7	3	0.66	0.46	14	6	0	0.85	0.26	15	5	0	0.88	0.22	0.36
L1HS21	0	3	17	0.08	0.14	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	17	0.00	0.00	0.04
L1HS26	5	4	9	0.39	0.49	8	1	3	0.71	0.43	11	2	2	0.80	0.33	11	4	3	0.72	0.41	0.42
L1HS32	9	8	2	0.68	0.44	13	5	1	0.82	0.31	15	5	0	0.88	0.22	13	4	1	0.83	0.29	0.32
L1HS34	0	10	10	0.25	0.38	3	14	3	0.50	0.51	1	10	6	0.35	0.47	1	5	12	0.19	0.32	0.42
L1HS39	11	3	1	0.83	0.29	15	1	0	0.97	0.06	12	0	0	1.00	0.00	11	1	3	0.77	0.37	0.18
L1HS43	4	10	6	0.45	0.51	8	11	1	0.68	0.45	12	7	1	0.78	0.36	7	9	1	0.68	0.45	0.44
L1HS44	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	19	0.00	0.00	0.00
L1HS46	16	3	0	0.92	0.15	16	0	0	1.00	0.00	20	0	0	1.00	0.00	13	0	0	1.00	0.00	0.04
L1HS57	0	3	17	0.08	0.14	0	2	18	0.05	0.10	0	3	17	0.08	0.14	6	4	9	0.42	0.50	0.22
L1HS73	19	1	0	0.98	0.05	20	0	0	1.00	0.00	20	0	0	1.00	0.00	18	0	0	1.00	0.00	0.01
L1HS74	0	1	19	0.03	0.05	2	5	13	0.23	0.36	2	7	11	0.28	0.41	1	5	12	0.19	0.32	0.28
L1HS77	6	12	2	0.60	0.49	19	1	0	0.98	0.05	18	2	0	0.95	0.10	17	2	1	0.90	0.18	0.21
L1HS78	1	6	13	0.20	0.33	5	3	11	0.34	0.46	3	4	13	0.25	0.38	0	5	12	0.15	0.26	0.36
L1HS85	0	0	9	0.00	0.00	0	3	17	0.08	0.14	0	2	18	0.05	0.10	0	2	14	0.06	0.12	0.09
L1HS86	14	0	0	1.00	0.00	14	1	0	0.97	0.07	12	1	2	0.83	0.29	17	1	0	0.97	0.06	0.10

(Table cont.)

L1HS104	7	8	5	0.55	0.51	9	5	4	0.64	0.47	5	12	3	0.55	0.51	10	5	3	0.69	0.44	0.48
L1HS110	20	0	0	1.00	0.00	19	1	0	0.98	0.05	20	0	0	1.00	0.00	18	2	0	0.95	0.10	0.04
L1HS112	0	2	17	0.05	0.10	0	5	14	0.13	0.23	1	4	15	0.15	0.26	1	1	7	0.17	0.29	0.22
L1HS117	8	1	1	0.85	0.27	9	3	1	0.81	0.46	9	8	1	0.72	0.41	7	4	3	0.64	0.48	0.40
L1HS118	0	6	13	0.16	0.27	3	8	8	0.37	0.48	0	7	13	0.18	0.30	0	3	15	0.08	0.16	0.30
L1HS131	10	0	2	0.83	0.29	8	3	3	0.68	0.45	5	3	4	0.54	0.52	14	2	0	0.71	0.44	0.42
L1HS132	2	12	6	0.40	0.49	4	13	2	0.55	0.51	3	8	9	0.35	0.47	0	9	11	0.23	0.36	0.46
L1HS153	6	6	8	0.45	0.51	2	9	8	0.34	0.41	4	7	8	0.39	0.49	3	6	8	0.35	0.47	0.47
L1HS157	17	0	0	1.00	0.00	17	1	0	0.97	0.06	18	1	0	0.97	0.05	18	0	0	1.00	0.00	0.03
L1HS158	4	12	4	0.50	0.51	9	7	1	0.74	0.40	6	13	1	0.63	0.48	2	14	4	0.45	0.51	0.48
L1HS160	18	0	0	1.00	0.00	18	0	0	1.00	0.00	19	1	0	0.98	0.05	16	0	0	1.00	0.00	0.01
L1HS163	4	11	4	0.50	0.51	1	13	6	0.38	0.48	12	6	0	0.83	0.29	5	9	5	0.50	0.51	0.45
L1HS166	0	3	17	0.08	0.14	4	7	9	0.38	0.48	3	10	7	0.40	0.49	1	5	12	0.19	0.32	0.36
L1HS169	13	1	1	0.90	0.19	8	8	2	0.67	0.46	12	4	1	0.82	0.30	12	0	0	1.00	0.00	0.24
L1HS171	3	9	8	0.38	0.48	0	6	13	0.16	0.27	1	15	3	0.45	0.51	1	2	10	0.15	0.27	0.38
L1HS172	14	4	2	0.80	0.33	5	12	3	0.55	0.51	12	5	3	0.73	0.41	10	9	1	0.73	0.41	0.41
L1HS173	15	1	0	0.97	0.06	17	0	0	1.00	0.00	12	0	0	1.00	0.00	4	1	3	0.56	0.53	0.15
L1HS177	20	0	0	1.00	0.00	18	0	0	1.00	0.00	19	1	0	0.98	0.05	12	0	0	1.00	0.00	0.01
L1HS178	17	3	0	0.93	0.14	19	0	0	1.00	0.00	19	1	0	0.98	0.05	12	1	0	0.96	0.08	0.07
L1HS180	1	6	13	0.20	0.33	1	9	10	0.28	0.41	4	10	6	0.45	0.51	4	8	7	0.42	0.50	0.44
L1HS197	11	1	1	0.88	0.21	8	1	0	0.94	0.11	12	0	1	0.92	0.15	14	0	0	1.00	0.00	0.12
L1HS213	20	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	18	2	0	0.95	0.10	0.02
L1HS214	20	0	0	1.00	0.00	17	0	0	1.00	0.00	19	0	0	1.00	0.00	17	3	0	0.93	0.14	0.04
L1HS220	0	0	20	0.00	0.00	1	1	18	0.08	0.14	0	2	18	0.05	0.10	0	4	16	0.10	0.18	0.10
L1HS222	1	6	8	0.27	0.40	0	3	16	0.08	0.15	0	1	18	0.03	0.05	0	2	18	0.05	0.10	0.18
L1HS226	0	3	17	0.08	0.14	0	1	18	0.03	0.05	2	6	12	0.25	0.38	1	4	15	0.15	0.26	0.21
L1HS228	17	0	0	1.00	0.00	14	0	0	1.00	0.00	18	0	0	1.00	0.00	12	1	1	0.89	0.20	0.05
L1HS231	20	0	0	1.00	0.00	17	2	0	0.95	0.10	20	0	0	1.00	0.00	18	1	1	0.93	0.14	0.06
L1HS233	1	4	14	0.16	0.27	1	6	11	0.22	0.36	1	7	11	0.24	0.37	0	7	13	0.18	0.30	0.32
L1HS235	1	15	3	0.45	0.51	1	9	7	0.32	0.45	1	11	8	0.33	0.45	3	12	5	0.45	0.51	0.48
L1HS242	4	11	5	0.53	0.39	0	11	8	0.29	0.42	2	11	7	0.38	0.48	4	5	10	0.34	0.46	0.44

(Table cont.)

L1HS260	20	0	0	1.00	0.00	19	0	0	1.00	0.00	18	2	0	0.95	0.10	19	0	0	1.00	0.00	0.02
L1HS287	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0.00
L1HS291	20	0	0	1.00	0.00	20	0	0	1.00	0.00	18	2	0	0.95	0.10	20	0	0	1.00	0.00	0.02
L1HS293	1	4	15	0.15	0.26	4	8	7	0.42	0.50	1	4	15	0.15	0.26	0	2	18	0.05	0.10	0.28
L1HS298	2	1	15	0.14	0.25	0	1	16	0.03	0.06	0	4	16	0.10	0.18	0	0	8	0.00	0.00	0.12
L1HS301	4	14	1	0.58	0.50	11	8	0	0.79	0.34	7	11	1	0.66	0.46	4	12	1	0.59	0.50	0.45
L1HS308	1	5	13	0.18	0.31	2	5	11	0.25	0.39	1	7	10	0.25	0.39	4	9	5	0.47	0.51	0.40
L1HS314	4	5	6	0.43	0.51	1	4	11	0.19	0.31	1	8	9	0.28	0.41	2	9	9	0.33	0.45	0.42
L1HS320	5	12	2	0.58	0.50	0	4	14	0.11	0.20	0	4	16	0.10	0.18	2	7	8	0.32	0.45	0.33
L1HS332	3	5	7	0.37	0.48	1	3	13	0.15	0.26	1	3	6	0.25	0.39	1	1	4	0.25	0.41	0.39
L1HS335	8	9	2	0.66	0.46	13	5	1	0.82	0.31	10	10	0	0.75	0.38	14	4	1	0.84	0.27	0.36
L1HS337	17	3	0	0.93	0.14	17	3	0	0.93	0.14	19	1	0	0.98	0.05	14	6	0	0.85	0.26	0.15
L1HS345	0	1	19	0.03	0.05	0	1	18	0.03	0.05	0	2	18	0.05	0.10	0	1	18	0.03	0.05	0.06
L1HS348	18	2	0	0.95	0.10	15	4	1	0.85	0.26	17	3	0	0.93	0.14	16	4	0	0.90	0.18	0.17
L1HS349	19	1	0	0.98	0.05	20	0	0	1.00	0.00	14	3	3	0.78	0.36	15	2	0	0.94	0.11	0.13
L1HS353	16	2	0	0.94	0.11	20	0	0	1.00	0.00	18	2	0	0.95	0.10	17	2	0	0.95	0.10	0.08
L1HS360	0	10	10	0.25	0.38	3	10	6	0.42	0.50	2	11	7	0.38	0.48	3	6	7	0.38	0.48	0.46
L1HS364	4	12	4	0.50	0.51	20	0	0	1.00	0.00	18	1	0	0.97	0.05	17	3	0	0.93	0.14	0.18
L1HS372	8	10	2	0.65	0.47	11	8	1	0.75	0.38	4	13	3	0.53	0.51	8	11	1	0.68	0.45	0.45
L1HS373	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0.00
L1HS375	6	12	1	0.63	0.48	11	8	0	0.79	0.34	4	16	0	0.60	0.49	11	9	0	0.78	0.36	0.42
L1HS378	18	2	0	0.95	0.10	8	10	2	0.65	0.47	14	3	3	0.78	0.36	13	5	1	0.82	0.31	0.31
L1HS391	18	0	0	1.00	0.00	19	1	0	0.98	0.05	20	0	0	1.00	0.00	19	0	0	1.00	0.00	0.01
L1HS393	1	2	14	0.12	0.21	0	0	19	0.00	0.00	0	0	19	0.00	0.00	0	0	14	0.00	0.00	0.05
L1HS395	7	9	1	0.68	0.45	8	9	3	0.63	0.48	3	12	5	0.45	0.51	9	7	3	0.66	0.46	0.48
L1HS404	1	9	10	0.28	0.41	0	0	18	0.00	0.00	0	0	20	0.00	0.00	0	2	16	0.06	0.11	0.13
L1HS406	17	3	0	0.93	0.14	16	4	0	0.90	0.18	18	2	0	0.95	0.10	16	4	0	0.90	0.18	0.15
L1HS410	0	10	10	0.25	0.38	5	10	5	0.50	0.51	3	10	6	0.42	0.50	7	11	1	0.66	0.46	0.47
L1HS413	0	11	9	0.28	0.41	1	9	9	0.29	0.42	0	7	13	0.18	0.30	3	6	10	0.32	0.44	0.39
L1HS415	17	1	0	0.97	0.06	18	2	0	0.95	0.10	18	0	0	1.00	0.00	20	0	0	1.00	0.00	0.04
L1HS418	4	10	6	0.45	0.51	13	4	1	0.83	0.29	5	12	3	0.55	0.51	2	8	8	0.33	0.46	0.44

(Table cont.)

L1HS423	18	2	0	0.95	0.10	17	0	0	1.00	0.00	17	1	1	0.92	0.15	15	1	1	0.91	0.17	0.10
L1HS426	1	14	5	0.40	0.49	7	5	5	0.56	0.51	2	5	9	0.28	0.42	3	6	10	0.32	0.44	0.47
L1HS427	5	13	2	0.58	0.50	15	5	0	0.88	0.22	8	9	3	0.63	0.48	11	8	0	0.79	0.34	0.39
L1HS430	0	2	18	0.05	0.10	0	4	14	0.11	0.20	0	0	20	0.00	0.00	1	0	19	0.05	0.10	0.10
L1HS437	1	14	5	0.40	0.49	0	3	17	0.08	0.14	1	4	15	0.15	0.26	2	10	7	0.37	0.48	0.34
L1HS442	10	10	0	0.75	0.38	17	1	0	0.97	0.06	14	6	0	0.85	0.26	8	7	2	0.68	0.45	0.29
L1HS446	0	2	18	0.05	0.10	0	2	17	0.05	0.10	1	6	12	0.21	0.34	0	0	17	0.00	0.00	0.14
L1HS447	12	7	1	0.78	0.36	11	3	3	0.74	0.40	14	5	1	0.83	0.30	13	4	2	0.79	0.34	0.35
L1HS448	9	2	7	0.56	0.51	3	13	2	0.53	0.51	14	5	1	0.83	0.30	7	8	2	0.65	0.47	0.45
L1HS450	12	4	4	0.70	0.43	20	0	0	1.00	0.00	19	0	1	0.95	0.10	18	1	1	0.93	0.14	0.17
L1HS461	0	3	14	0.09	0.17	0	1	19	0.03	0.05	0	1	18	0.03	0.05	0	0	17	0.00	0.00	0.07
L1HS480	3	8	9	0.35	0.47	4	8	6	0.44	0.51	5	10	5	0.50	0.51	4	10	6	0.45	0.51	0.50
L1HS486	3	7	10	0.33	0.45	7	9	4	0.58	0.50	1	2	17	0.10	0.18	0	1	18	0.03	0.05	0.30
L1HS493	5	8	6	0.47	0.51	5	8	7	0.45	0.51	9	7	3	0.66	0.46	9	2	4	0.67	0.46	0.49
L1HS508	16	4	0	0.90	0.18	17	3	0	0.93	0.14	11	8	1	0.75	0.38	17	2	0	0.95	0.10	0.20
L1HS512	19	1	0	0.98	0.05	18	0	0	1.00	0.00	19	0	0	1.00	0.00	17	0	0	1.00	0.00	0.01
L1HS513	0	4	16	0.10	0.18	6	10	3	0.58	0.50	4	7	9	0.38	0.48	2	6	10	0.28	0.41	0.39
L1HS516	2	8	9	0.32	0.44	1	2	16	0.11	0.19	6	9	5	0.53	0.51	3	7	6	0.41	0.50	0.41
L1HS526	5	13	2	0.58	0.50	13	6	0	0.84	0.27	3	12	4	0.47	0.51	3	7	9	0.34	0.46	0.44
L1HS552	0	6	11	0.18	0.30	5	7	8	0.43	0.50	2	14	3	0.47	0.51	1	5	12	0.19	0.32	0.41
L1HS558	16	4	0	0.90	0.18	16	3	1	0.88	0.22	17	3	0	0.93	0.14	18	2	0	0.95	0.10	0.16
L1HS561	0	1	19	0.03	0.05	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0.01

(Table cont.)

Table 2B - L1Hs Ta X-linked associated human genomic diversity.

	African American							Asian/Alaskan Native[d]						
	Genotypes					Population		Genotypes					Population	
	Females			Males				Females			Males			
Element	+/+	+/-	-/-	+	-	f[c]	H[a]	+/+	+/-	-/-	+	-	f[c]	H[a]
L1HS24	1	5	3	1	8	0.30	0.40	3	2	1	8	2	0.73	0.43
L1HS28	5	4	0	6	3	0.74	0.42	0	3	3	3	7	0.27	0.44
L1HS30	0	5	5	4	5	0.31	0.48	1	4	2	7	3	0.54	0.53
L1HS125	7	1	1	7	1	0.85	0.26	6	0	0	10	0	1.00	0.00
L1HS562	1	5	3	1	8	0.30	0.40	3	2	1	8	2	0.73	0.43
L1HS564	0	3	7	2	7	0.17	0.32	0	0	6	1	9	0.05	0.10

	European/German Caucasian							Egyptian							
	Genotypes					Population		Genotypes					Population		
	Females			Males				Females			Males				
Element	+/+	+/-	-/-	+	-	f[c]	H[a]	+/+	+/-	-/-	+	-	f[c]	H[a]	AH[b]
L1HS24	5	3	1	7	3	0.71	0.44	5	8	4	1	2	0.51	0.50	0.44
L1HS28	1	5	3	9	1	0.57	0.49	9	6	1	2	1	0.74	0.43	0.44
L1HS30	2	4	3	6	4	0.50	0.53	3	10	3	3	0	0.54	0.39	0.48
L1HS125	9	0	0	9	0	1.00	0.00	16	0	0	3	0	1.00	0.00	0.07
L1HS562	5	3	1	7	3	0.71	0.44	5	8	4	1	2	0.51	0.50	0.44
L1HS564	0	2	7	1	9	0.11	0.20	0	3	13	0	3	0.09	0.09	0.18

a. This is unbiased heterozygosity. H=(2*sample size * (1-sum freq of homozygotes))/(2*sample size – 1)
b. Average heterozygosity is the average heterozygosity for all populations.
c. f represents the frequency of the element.
d. Asian and Alaskan Native samples were used interchangeably as a geographically unique human population

Table 3 - PreTa L1 primers, PCR Conditions, and associated human genomic diversity

Name	Accession	Chrm. Loc.[1]	Forward Primer	Reverse Primer	Human Diversity[2]	AT (F,R)[3]	AT (ACG)[3]	PCR Product Sizes[4]		
								Filled	Empty	Subfamily Specific
L1AD1	AC080166.6	2	AATTCGCTGCATAATTTCTT	AAACATATGGCCATCTTGAC	FP	55	60	6835	249	578
L1AD2	AC090955.2	3	TTTTCTCCATGACTTGAGATGGT	TGCAATCATGAAAACCAGTG	FP	60	60	6308	245	265
L1AD3	AC018878.8	2	TGCACATGGATGTGTAAGAATAC	TTCTTCCCATAAGCATTGGT	FP	60	60	6448	339	245
L1AD4	AC053545.5	4	TTGATGCATTTCTGCATAAGG	CCAAGATTTTGGCTAGCATTT	FP	55	60	4528	295	188
L1AD5	AC079801.2	16	TCATCTCACAGAGCTCACAG	CTAGGAATCCTTCTGTCTGG	NP		60	749	326	150
L1AD6	AC073647.9	7	GCAAACACTGGTTCAAGAAG	TGGAGATAGTGTAGGCACAG	FP	55	60	1741	87	233
L1AD7	AC093607.3	4	INSERTED IN REPEATS		R					
L1AD8	AC079926.7	4	GCCTCTTTCTTAGTCAAGCA	AGGTCACAAGGGACATTTCT	NP		60	857	417	208
L1AD9	AC012593.8	2	CAGGTAGGGGAAAGGAGGAG	TGGGCTTATTATCCCCTTGA	FP	55	60	1034	392	342
L1AD10	AC016906.7	2	TGTATTTACCGGGGATGAGG	GCTGTCCCAAATTTCCAGAG	IF	60	60	3602	172	229
L1AD11	AC018465.8	2	GCACCTTGCTATTTGTTTTCT	CCCTAGAGCAATCACCAAAGA	FP	60	60	6515	458	185
L1AD12	AC083950.4	2	GGATAGGCAATGTGTTAGGT	TGCAGAGGCAGTTGTAACAT	FP	55	60	1106	603	303
L1AD13	AC097484.3	4	AAACCTATACATAGAAAATTGCTG	ACCCAGAACAAATGAACACT	FP	60	55	1368	473	424
L1AD14	AC012665.8	2	TTCTGCAACTATAGCCGTAA	ACAACAGACACAGAAGCAAA	IF	60	55	6187	136	173
L1AD15	AC093584.3	4	INSERTED IN REPEATS		R					
L1AD16	NG_000004.1	UNK	GGTTGAGAACCACTGTCATAA	GCCAGTGCTTAGATTTACCA	FP	60	60	6213	145	260
L1AD17	AC105459.1	7	ATTCCCCATTTTACGATTTT	GCTACTGCCGTGTTTTACA	FP	55	60	440	276	309
L1AD18	AC096764.3	2	AGATGCCCGGTCTACTACTT	AGCACTTTAAAGGCATCAAC	FP	55	60	3467	151	249
L1AD19	AC009156.9	16	ATATTGGCCAAAGCCTCTTA	TGGCAAGTCCTGAATGATAA	IF	55	55	3974	88	191
L1AD20	AC009156.9	16	CATTAGCAAGCTGATTCAAA	CTTTTGCCATGATTAGTGGT	HF	55	60	474	147	205
L1AD21	AC097522.4	4	CAGAAAGTCATCTCATCTTCC	TAAAGCATTCGTTGTTGTTG	FP	55	60	6528	353	587
L1AD22	AC092570.3	2	CCTCCTCACCTCCTTTTAAT	ATGAAGGGAAACGAGAAAAG	FP	55	60	562	63	220
L1AD23	AC018673.4	12	INSERTED IN REPEATS		R					
L1AD24	AC097451.2	4	TCGTTCCTCATCTCTTTGTT	AGCAAAAGCAGTCACTTTTC	FP	55	55	3467	382	396
L1AD25	AC023154.5	4	INSERTED IN REPEATS		R					
L1AD26	AC096769.3	4	TTGAGTTTTCCCTCCATGAAA	TCTGATGAATTGTGCCTGACA	FP	60	60	381	157	263
L1AD27	AC093877.3	4	AATATTTAACATGGCCCATAA	GGCATTGGTGTCAATGAGAA	FP	60	60	1171	110	834
L1AD28	AC096749.2	4	GAAGGCTTTATACTCCTTCTTGGA	TCATGGGAGATTTTTCAACTTTC	FP	55	60	6459	419	330

(Table cont.)

L1AD29	AC105150.2	8	GGACAGAAATACTGGCATCT	CACAATCTTATCTCAAGGGAAT	FP	60	60	6398	318	354
L1AD30	AC055820.7	18	CTTGATGGCAATACAGCCTAA	CCATTAATGTGGGCTCATAATCT	FP	60	60	1855	78	208
L1AD31	AC018626.8	18	GGGAAACGACAGAAGATGGA	GAATTTTGATTTGTGGGCATA	FP	60	60	1143	209	204
L1AD32	AC091613.3	1	END OF CONTIG		EC					
L1AD33	AC092798.3	3	INSERTED IN REPEATS		R					
L1AD34	AC012642.5	5	GGCTTGTGCTACACAGAGTT	CCAACCAGGAACAATAAAAG	FP	55	55	2816	519	247
L1AD35	AC021538.8	UNK	AAATGCCCACAAAATTCCTG	CCATGGGAGCTACTGGAAAA	FP	55	60	984	386	479
L1AD36	XM_037013.1	UNK	END OF CONTIG		EC					
L1AD37	AC099515.2	5	INSERTED IN REPEATS		R					
L1AD38	AC026703.4	5	CCCAGTTCTCCAAAATATCA	CACTTGCCTATGGTTCATTT	FP	55	55	5984	240	468
L1AD39	AC078857.12	3	INSERTED IN REPEATS		R					
L1AD40	AC078857.12	3	TCGTGACCTTATTAGCCACT	CCTCCATTTGCTACCTAGAG	FP	60	60	1680	512	633
L1AD41	AC078857.12	3	TGTTATTTCAGCTTTAACCATCAA	TTTAAAAATCAAGTATGGGAAAAA	FP	55	55	1202	141	242
L1AD42	AC093515.3	16	INSERTED IN REPEATS		R					
L1AD43	AC011597.27	3	INSERTED IN REPEATS		R					
L1AD44	AC079943.18	3	ATGCCATCCCCTGGATTT	TGGTTGCTCCAAAGGAACTT	NP		60	6591	530	316
L1AD45	AC061710.16	3	GAGCAAATTTGTCAGACAGAACA	TGGGATGGTTGAAATCAAATG	FP	60	60	3854	147	199
L1AD46	AC072051.8	UNK	CCCTATTTTCCCCATCATCA	AAGCAGGCAGATGGTCACTT	FP	55	60	3706	69	166
L1AD47	AC008006.10	18	CGTCACACACATAACCAGAG	GATCAGGAATATGGCAAAGA	FP	55	55	471	212	284
L1AD48	AC027553.6	UNK	TGCATGAAGCACTACTCAAAGA	TGCAAGATGTGTCAGTATTTAGC	FP	60	60	6181	106	226
L1AD49	AC018991.10	UNK	INSERTED IN REPEATS		R					
L1AD50	AC008948.8	5	INSERTED IN REPEATS		R					
L1AD51	AC008728.7	5	INSERTED IN REPEATS		R					
L1AD52	AC093566.3	8	INSERTED IN REPEATS		R					
L1AD53	AC020783.8	8	INSERTED IN REPEATS		R					
L1AD54	AC068062.5	10	CCTTTGTTTCTTGGGTGTGG	CCCACATCACCAAACCATTT	FP	60	60	357	128	212
L1AD55	AC064875.5	2	GCCACACTCCTTTGTTTGCT	CAAGCACAAAAGCAGGAACA	FP	60	60	724	193	273
L1AD56	AC073275.8	7	INSERTED IN REPEATS		R					
L1AD57	AC010747.10	2	CGGAAAATTTGTTACTTGCT	AGGTATGCTGCATTTCTTTC	FP	55	55	3903	97	272
L1AD58	AC012509.13	2	CCCTGGATGCTGAGTTTCTT	TCCATCTGGCATTGACTCAG	FP	60	60	1062	139	213
L1AD59	AC009964.11	2	TGGGACATTGACTCCTACTC	GGCATAGGTTTCTGGAAGTA	NP		60	760	340	282

(Table cont.)

L1AD60	AC009961.11	2	INSERTED IN REPEATS		R					
L1AD61	AC078851.4	2	TTTATGCTGATCACTGTTCTTC	AACTAGTTGCATCGTGATCATA	FP	60	55	2090	80	208
L1AD62	AC016720.9	2	CTTTCGCATCATCGTAAAGT	ATTGCCAACTGGTTACAAAG	FP	55	55	2886	114	261
L1AD63	AC012492.9	2	AAAAACCCTTTAAGCTCAGT	TGGAAGCATACAAAATGAAA	FP	55	55	6402	342	180
L1AD64	AC069285.8	7	GCCACTGCTAATCAATTCAC	CCAAAGCAGACACAATTTCT	PARALOG	55	60	6131	77	172
L1AD65	AC026029.8	4	TTTCCTCAAAGTTGATGCTC	CCTGGAAGGCATAACTGATA	NP		55	6787	271	575
L1AD66	AC025223.6	2	TATCCAAATATCCCTTGCAG	TTGTAGTTTGTGGAACTGGA	PARALOG	55	55	717	201	197
L1AD67	AC095347.6	12	INSERTED IN REPEATS		R					
L1AD68	AC069242.13	3	CCTATGGATGAAAAATGGAC	TCTGAAAATGTTGCCATTG	FP	55	55	294	111	176
L1AD69	AC092325.2	16	INSERTED IN REPEATS		R					
L1AD70	AC079841.10	3	TCCAAGAGCAGGCAGTATTA	TTCCTGACTACTCCAGTTCAG	FP					
L1AD72	AC092468.9	3	GTGCAGGTGTAAGGAAGAAA	GTCTTCAAACCAGACTGCAT	FP	55	55	546	93	218
L1AD73	AC097657.3	4	TGATTTGCAGTATTTTTCCTT	GCATGACCCAGATTAGAAAA	FP	55	55	1148	126	168
L1AD74	AC097463.2	2	NO RESULTS		NR					
L1AD75	AC092018.2	1	TTTCTCTCCCTCAAGCCTTTT	CCAAAATTCATGCTGGGAAC	IF	60	60	1636	129	124
L1AD76	AC027345.5	4	AAACCTCCCTTTAGTCTCCA	CACCAGACCCAATTTTTAGA	FP	55	60	4500	221	173
L1AD77	AC097110.1	4	TCAAGGAAGGGAGTTAAAAA	ACTTCTTTCATGCCCCTTAT	HF	55	55	991	729	237
L1AD78	AC026439.4	5	TCTTGAGGCTTGCAAATACT	ATGAGCAACAAGAAATCACC	FP	55	60	1559	295	306
L1AD79	AC016620.7	5	INSERTED IN REPEATS		R					
L1AD80	AC092185.3	3	AAGCAGTATGTCTGGCACA	ACAAACTGACACTCCAAACC	FP	60	55	6148	72	197
L1AD82	AC022165.8	16	GGTGTCTCCACAGTTGATTC	CCACCGCCAGATTTTACTA	HF	55	55	2876	117	196
L1AD84	AC090525.8	12	TTCCCTGGGTCACTTTTCTC	TGCCAAATTGCTTTGCATAC	FP	55	55	2068	255	333
L1AD85	AC026120.33	12	INSERTED IN REPEATS		R					
L1AD86	AC093865.2	2	ACATGATGTCCCATCTTCCA	AAGAGCCATATGAGAGCTTCC	FP	60	60	1046	271	304
L1AD87	AC022446.6	5	AATTTTTCCCCACATGTTC	ACAGAATGGATTTAGCTTGC	FP	60	60	3761	118	248
L1AD88	AC090519.3	15	INSERTED IN REPEATS		R					
L1AD89	AC084819.17	12	INSERTED IN REPEATS		R					
L1AD90	AC092601.3	2	INSERTED IN REPEATS		R					
L1AD91	AC008571.6|	5	TGCTAAACAGAAGGCACATA	ATAGATCCATCTGCCAAATC	FP	55	60	6266	170	314
L1AD92	AC092638.2	2	TTATCCAAAGAAGGGGAAAGG	TTTGCCTTATAAGCATTGTGAAAA	FP	55	55	6224	181	195
L1AD93	AC096653.1	4	CAACACTCATTACAACCTGTG	CAGAGTTTATCAGCCAGACC	FP	60	60	2336	382	399

(Table cont.)

L1AD94	AC092581.2	4	CTCCACGTTAACAGATAGGG	TGAGCTTCACTTAACCACTG	FP	60	60	507	341	239
L1AD95	AC096569.1	2	CCAGCACTGATTTCATAGATGC	TTCAGACAACTGAAGTGCCTTT	FP	55	55	6161	89	224
L1AD96	AC092631.1	4	TAATTAGGTAACGCCTGTGG	CAGGAAGCCTAAACTGCTT	IF	60	60	932	98	245
L1AD97	AC008709.6	5	CCCCAGGCTTTTGAAAATTA	ATTCTCGGGGTCCCAATTAC	FP	60	55	6164	111	214
L1AD98	AC060796.7	17	ATGGAAAGGGGAAGATTTTA	GGCTATACTACAACATCCCTCA	FP	55	55	6164	126	203
L1AD99	AC090791.6	11	GTGACACAAAAAGCACAATTAC	CAATGATTCATGAGTTGGAA	FP	55	55	2737	292	303
L1AD100	AC026729.5	5	CCTGGGTCACAATATGAAGA	TCTGATAACCAGAAGATGAAGA	HF	55	60	6324	258	352
L1AD101	AC025467.5	5	AGTCTCCCTTTCAGAAGCA	AATGCTGGGAATCTTACCTC	IF	55	55	6091	66	163
L1AD102	AC025467.5	5	GAATGGGGTGTGCTGTAA	TTTTAACAAGATCCCAGACC	IF	60	55	3721	78	164
L1AD105	AC010275.6	5	ATTCTCGGGGTCCCAATTAC	CCCCAGGCTTTTGAAAATTA	FP	55	55	6164	111	214
L1AD108	AC008550.5	5	CACAATCATACCTTCCCAACTG	CAGATGAGACTTTGGACGTGA	FP	60	60	6154	84	187
L1AD110	AC092721.2	16	ATTTTGTGGTTCAGCATTTT	CATAGAAAAGGGAACAAATGA	FP	60	60	1590	82	226
L1AD111	AC092357.2	16	AAAAGTTGTTTTCCTGATTTTT	AGTTTTCTCTGCAGCTCATC	FP	56	55	6252	188	184
L1AD112	AC034219.5	5	TTTCCAAAAACAGCTAGGAG	CGTTTTTCTAGCTTAGCAATG	FP	55	55	406	106	209
L1AD113	AC005406.2	UNK	ACCTTGATTGCAAATTGTTT	GGTTTCTTGGCCTCTTTACT	FP	60	60	2881	80	189
L1AD114	AC020651.19	3	INSERTED IN REPEATS		R					
L1AD115	AC084032.23	12	AACTGCCATGAAAACTTACC	AAAGATTGTCCACATCAAGG	FP	55	60	253	100	190
L1AD116	AC025176.5	5	END OF CONTIG		EC					
L1AD117	AC022024.6	10	CAGCAACCATAGGTTGATAAG	GGATTACTGCCCAAAGAAAC	FP	60	60	852	487	310
L1AD118	AC026113.25	12	GACTGCTGGATCAAATGTTAG	ACCACCTTACTCCTGCTACA	R	55	60	6231	188	272
L1AD119	AC024941.30	12	CTTTATTCATGGCAGAAAGC	CTCATGAGATCTGGTTGTTT	R	55	60	1347	112	249
L1AD120	AC066613.7	UNK	INSERTED IN REPEATS		R					
L1AD121	AC010857.8	4	INSERTED IN REPEATS		R					
L1AD122	AC011712.6	18	CCCAGGGGAATATATGGAAATTA	AATTGAATGCAGATGGTTTTACC	FP	60	55	6631	139	608
L1AD123	AC010928.7	18	CCAGGAGTCAGAGGATTACA	TCTGTTGTGAGAAGCAAATG	FP	60	60	410	98	172
L1AD124	AC013759.6	18	INSERTED IN REPEATS		R					
L1AD125	AC013759.6	18	AAACGGTGAAGGAAATGTTG	GACATGAGCAACCATCAGGA	IF	60	60	513	236	309
L1AD126	AC021082.4	5	INSERTED IN REPEATS		R					
L1AD127	AC012323.7	16	INSERTED IN REPEATS		R					
L1AD128	AC025097.41	UNK	INSERTED IN REPEATS		R					
L1AD130	AC039057.8	UNK	INSERTED IN REPEATS		R					

(Table cont.)

L1AD131	AC073258.9	7	INSERTED IN REPEATS		R					
L1AD132	AC017014.4	2	GGGAAGTGAAGGCTAACATA	ACCATGGAGCTCAATTTACA	FP	60	60	469	84	187
L1AD133	AC069294.5	7	GGTTGAGAACCACTGTCATAA	GCCAGTGCTTAGATTTACCA	FP	60	60	6212	145	259
L1AD134	AC084732.1	4	CTACCCAGAACAAATGAACAC	AACCTATACGTAGAAAATTGCTG	FP	60	60	1368	475	422
L1AD135	AC008276.4	2	CTCAAGGGTTCTCATCACTAA	GGAAAGGATACCACAATCAA	HF	60	60	1871	87	191
L1AD136	AC017015.4	3	TGGCTGACAAATTGGTGATT	CCCATGTGAACTGCATTGAA	FP	60	60	712	293	217
L1AD137	AC010970.3	Y	INSERTED IN REPEATS		R					
L1AD138	AC012284.5	15	GAGCTGAAGAAACAAAGGAA	ACCTCAAATTCATTTTGGAA	FP	55	60	780	75	200
L1AD139	AC009479.4	Y	INSERTED IN REPEATS		R					
L1AD140	AC010722.2	Y	TTCAGGAACATTGCTATGAGGAT	TAGGCATTTATCATGTGCTC	FP	55	55	1643	218	283
L1AD143	AC079175.24	X	CAGTAAACTGGGCTGCTATC	GAGAGTCAAGCAGTGGGTAA	FP	55	60	5078	80	208
L1AD144	AC023842.5	8	CACAAGATTCAATACCTGAGTGACA	TGGGCATTACTAGTTGAACCTAAAG	FP			1641	141	261
L1AD145	AC087883.12	12	GAAGGAAGCCCCCATATGAT	GAGGTGAAAGGCCATTAAAGAA	FP	60	55	473	147	243
L1AD146	AF280107.1	UNK	END OF CONTIG		EC					
L1AD147	AC063951.22	12	END OF CONTIG		EC					
L1AD148	AC024060.5	3	AACTTCCTTAGGACCTCATTT	TGTGTTTAACGTTCTAAACCTG	FP	60	60	1361	65	229
L1AD149	AC087433.4	15	CCGAAACACAGATAAGCACT	AGTGTAAAAATCTGCATAGCC	FP	55	55	2160	508	274
L1AD150	AC073572.19	12	ATTCCCCCAATTCTCCAAAA	GCAAGGGCCAACTATGCTAA	FP	55	55	1195	124	187
L1AD151	AC023795.18	12	INSERTED IN REPEATS		R					
L1AD152	AC079865.14	12	GGGAGATCCAGACATACAAC	TGTGTAACTCTTTTGCGATG	FP	60	60	569	369	341
L1AD153	AC058784.17	13	INSERTED IN REPEATS		R					
L1AD154	AC023812.7	3	ACCTCTACCTTACCACACCA	CCTAACTCAGGTCATTCTGC	FP	60	60	1475	175	260
L1AD155	AC018923.21	3	INSERTED IN REPEATS		R					
L1AD156	AC008436.5	5	INSERTED IN REPEATS		R					
L1AD159	AC008496.5	5	INSERTED IN REPEATS		R					
L1AD160	AC034194.4	3	AGAGCTACATGGCTAAATGC	TCTGCAGTTTTAACACCTCTT	IF	60	55	543	238	261
L1AD161	AC011546.6	19	INSERTED IN REPEATS		R					
L1AD162	AC020717.3	X	TTCCTATAGGCTTGAATGGA	TTTTGGTGCCCAATAGTATC	FP	55	60	2923	198	219
L1AD163	AC007132.3	2	CCCAGTATGTCCCTCACTCAG	TAGGCAAACCCCAATTGAAA	FP			6359	315	351
L1AD164	AC006968.2	X	TTCCCTGTCCAATGTAAAGAA	AAAGTGCATATTGCACAGGA	FP	55	55	836	107	158
L1AD165	AC010685.3	Y	INSERTED IN REPEATS		R					

(Table cont.)

L1AD166	AC010889.3	Y	CCCTAACATTTCAAAATGCACTG	ATTTTTCCAACTACTGGCACTCA	FP	60	60	1256	162	214
L1AD167	AC006334.3	7	INSERTED IN REPEATS		R					
L1AD168	AC009489.3	Y	TGCCTTTATAATATGGAAATGCAG	TGCTCATGGAGTCAGAATATGAA	FP	55	55	1080	196	183
L1AD169	AC011745.4	Y	TCCCATTGCATTTAGCAGATT	AGGCCTGTATTTCAATTGTGCTT	FP	60	55	3676	95	265
L1AD170	AC007278.3	2	GTCTATTAATCCCCCTCCAC	CAACGTTGAAAAGATGTAGAGA	FP	60	52	6149	87	174
L1AD171	AC006992.2	7	TGGAACTATTTCAGGAAATTAAA	AACAAGGGGGAAGAGAATAA	FP	55	55	6278	197	234
L1AD172	AC006362.2	7	INSERTED IN REPEATS		R					
L1AD173	AC015542.17	3	TTCCAATATACTTTGCCCTTA	AGTAGGCATCAGCAACAGTC	FP	55	55	546	393	322
L1AD174	AC022013.3	3	TTTGGGGAGAACTATCTGTG	GCTTGGACATTGGAATTTT	FP	60	54	399	118	188
L1AD176	AC026204.4	3	GCACTCTCATTTACTGCTGA	CCACCTTTTACTATTTTGGTG	IF	60	55	838	494	195
L1AD177	AC018514.7	14	ACCAGATGGAAGCTAGATGA	AAGTTTCCAAGGGAAATCAG	FP	60	55	6370	256	373
L1AD178	AC058791.3	7	ATTGTTTAGGGGAAAAGGAC	CCAAAAGCAGGTTAATTCTC	FP	55	55	629	203	322
L1AD179	AC013738.4	10	ACTCCACTTTAATTCGCAAG	GAAGGCGAGAAACTGTAGAA	FP	55	60	1056	113	289
L1AD180	AL627250.8	X	INSERTED IN REPEATS		R					
L1AD181	AL449304.19	9	TTCCATAGCCATTGATTACA	AATTTTCAGGCACGTTTTTA	FP	55	55	652	286	446
L1AD182	AL137787.11	X	INSERTED IN REPEATS		R					
L1AD183	AL445312.5	X	GTCCAGAAGTCTCTCCTGTT	CGATTGCAGGCTTTCTAATA	FP	55	60	2873	105	413
L1AD184	AL360020.15	9	INSERTED IN REPEATS		R					
L1AD185	AL391260.13	10	TTCTGTAGGGCTCCTGACTA	ATTCACAGTTCCCCGTAGTA	FP	60	55	7905	185	1829
L1AD186	AC016951.9	3	ACTTGAAATTGGGGTAGATG	ATTTTCTAGAGGGCTCCTTG	IF	60	59	843	190	206
L1AD187	AL365258.24	1	INSERTED IN REPEATS		R					
L1AD188	AL603765.6	1	INSERTED IN REPEATS		R					
L1AD189	AL596326.5	1	TGTTTCATGGAGTGTATTTCA	TGCAATGTTAGAAGAAGTGG	HF	55	55	456	198	289
L1AD190	AL606752.11	1	GCTTGACACATAGTGCTTGA	AAATGTGGCATTATTTTCACT	FP	60	60	462	250	193
L1AD191	AL589877.13	X	ACCCAGAAACGCATATACAC	GCAAATTGCAACAAGATAAA	FP	55	55	1926	591	352
L1AD192	AL513493.11	1	TGTCCAATTAAAAGGCACAT	TGGAATATCTTTTTCTGCCTA	FP	55	60	941	134	322
L1AD193	AL359733.15	1	TCTTTTACTCCCAAAAGGAA	TTGGGTAGATGAAGATGACC	NP		55	1900	260	292
L1AD194	AL357873.17	1	GCCCTGGATGTAGTGTATGT	CTCTCTCTTCATCCGTTCAG	FP	55	55	974	144	256
L1AD195	AL592494.4	1	NO RESULTS		NR	55	55			
L1AD196	Z82209.2	X	TTCTCTCCTAACCCTCTTGG	TTTAGGGTATGCGGTAGAAG	FP	60	55	6581	349	385
L1AD197	AL354949.10	1	GAAACTGAGATTCACGGAAG	AGTTTCTCATCCCACCTTCT	FP	60	60	6437	360	467

(Table cont.)

L1AD198	AL138785.8	1	GCTTCACCTCACTAGCCTTA	CTCACAAAGCAGCATTTACA	FP	60	60	456	87	163
L1AD199	AL445197.4	1	TTCAGCATATCTGCAAAGTG	GAAAGGATTCTCATTTCCTG	FP	55	60	626	216	341
L1AD200	AL136224.24	6	CAGTCTATCAATTCCTGTTGG	TGATCATCCAGCTCAATTACT	FP	60	55	2353	472	440
L1AD201	AL607144.5	13	CAGACTTGGGCATCTTTTAG	AAAACATCAGGGCCAAATA	FP	55	57	1328	148	178
L1AD202	AL513324.8	10	INSERTED IN REPEATS		R					
L1AD203	AL390834.24	10	INSERTED IN REPEATS		R					
L1AD204	AF245226.1	21	INSERTED IN REPEATS		R					
L1AD205	AL596342.3	1	GACTCTTCCCCTTGAGAATC	GCATGCCTACGATCTCTTAT	FP	55	55	381	222	253
L1AD206	AL603902.4	6	INSERTED IN REPEATS		R					
L1AD207	AL592067.4	13	ATTTAGGTATGCGTTTCAGC	ACATCTCTTCATGCCTTCAG	FP	55	55	999	422	238
L1AD208	AL353743.22	9	ATCTCCTATCCCCTTAGCTG	AACCCAAGAGTCACAGTTGA	HF	60	60	1978	530	280
L1AD209	AL139282.10	1	TTGAGTCAAGGAAAAATAATGA	AAAGCAAGGCAGGTATGTTA	FP	60	60	1667	214	245
L1AD210	AL512504.9	X	INSERTED IN REPEATS		R					
L1AD211	AL590439.12	10	ATATTGATTTGGCATCCTGA	GTAAACGTTCTAGCCAAAGC	FP	60	60	6207	155	169
L1AD212	AC007347.3	16	CACGGGAGAAGATTTATGTC	TTGTACCTACTCCACCCAAG	FP	54	55	6400	210	310
L1AD213	AC007262.4	14	GCCATAAACAGAAAACCATT	GTTGCAGAAATAACAGCACA	IF	60	60	494	182	294
L1AD214	AC007221.2	16	GCAGTCAACATCTTCCAGTA	TGAGCTAGAATCCCAAAGAT	FP	55	60	6267	135	324
L1AD215	AC007115.1	12	TGAAGAACCTTCACGTAAGAA	AAATATGATGCTTTGCTTCC	FP	60	55	556	176	362
L1AD216	AC006143.1	X	GAGGCTTACTGGAAGCATAG	CTCACGGTTGATGTCACTTT	FP	60	60	1494	430	520
L1AD217	AC011594.8	12	CTGGCCAAAGAGGTAGTTT	CAAAAGAGCATGGTACTGGT	NP		55	7620	479	537
L1AD218	AC004141.1	7	TCCTTAACCTAGTTGCTCCA	AGGGTACATTGAAGTTGAGG	NR	60	60	624	340	458
L1AD219	AC002076.1	7	AGGGAATATTTGGGACATCT	CCCCACCACACTAGAAACTA	IF	60	60	6418	354	391
L1AD220	AC003085.1	7	CCAGGGAACTTGATTTTAGA	CAATTGGATAAGAGGGACTG	FP	60	55	6500	303	199
L1AD221	AC004161.1	UNK	INSERTED IN REPEATS		R					
L1AD222	AC006204.1	7	TTTGGAAGCTTCACTTTAGC	TGGCCTTAATATTTTAGCAAC	FP	60	60	590	167	246
L1AD224	AL356096.11	13	INSERTED IN REPEATS		R					
L1AD225	AL513355.16	10	CGGTTCTAAACACCATTTGT	TTATGGCCCTTAATTTCATC	FP	60	60	1739	177	192
L1AD226	AL358873.25	6	GCATCTTTGAATCAACAAGTC	TGTATCTAACTATTCCCAGTGATT	FP	60	55	986	238	751
L1AD227	AC004822.1	X	TTGAGAGCATCCATATTTCC	CCAACCTCAGATTACCAAGA	FP	55	60	768	115	202
L1AD228	AC005053.1	7	INSERTED IN REPEATS		R					
L1AD229	AL450312.10	9	INSERTED IN REPEATS		R					

(Table cont.)

L1AD230*	AL583806.7	6	GCAATCCATAGACAACCAAT	AGGAGGAATATGCAAACTGA	HF	55	55	2249	599	338
L1AD232	AL583825.8	1	TCCCAGAACTACCTCATAACA	GAGGAAGACAGTGTCACAGA	IF	60	60	1162	219	329
L1AD233	AF207955.1	21	AGGGGTAGATTTTGTTCAGA	AGGACCATTTGCAATGTTAG	FP	60	60	1283	747	667
L1AD234	AL391992.8	10	TGGCTAGTCACCCTAAAAGA	GTTTTATAGGCTTGCATTGG	FP	60	55	6487	388	360
L1AD235	AL160234.3	14	GGAGCTATTAAGCCACAAAA	GAGAGGGTATCCTCGTCTTA	FP	55	60	6771	694	326
L1AD236	AL079307.7	14	GAATGGGGAATTATACGTGA	GTAAGGCACTTGGAAATGTG	FP	60	60	6260	196	295
L1AD237	AL162431.17	1	AAGTGAATGTGGATTTACCC	TCTCAAGGAAATCAGCTCTT	FP	60	60	6499	435	324
L1AD238	AL389895.3	14	ACTTTTATGCCTGAAACCTG	ATCCTTTCTCAGAGGGATCT	FP	60	60	6370	325	278
L1AD239	AL357045.10	1	INSERTED IN REPEATS		R					
L1AD240	AL591770.1	14	GTCTCAGACACACAAGCTCA	TTGGCCACTCATCTATCTTT	HF	60	60	540	222	258
L1AD241	AL512310.3	14	INSERTED IN REPEATS		R					
L1AD242	AL136960.4	13	CCCCTGAAGAGTCCATATAA	CCTAACAGTCAGGAAAGCTG	FP	55	55	6347	288	197
L1AD243	AL445466.9	1	CTGCTTGTCTTTGGTCTGAT	GTGATCCTGTAGGCCTTCTT	FP	60	60	2933	410	1229
L1AD244	AL512790.1	14	GCATCCGTTTCTCTGATG	TGCAGATTGTACAGAAAAGC	FP	60	60	1394	166	296
L1AD245	AL136295.3	14	ACTTTAGGATTCCGTGGTTT	AATGCTGTTAGAGGAGGATTC	FP	55	60	2193	482	222
L1AD246	AL391838.9	13	INSERTED IN REPEATS		R					
L1AD247	AL512662.8	UNK	INSERTED IN REPEATS		R					
L1AD248	AL138694.18	UNK	INSERTED IN REPEATS		R					
L1AD249	AL133241.3	14	INSERTED IN REPEATS		R					
L1AD250	AL121852.3	14	CCCTCAAGAACGATTTTATG	TGTCTAGAATGTTCCCCTTT	FP	60	60	6397	280	237
L1AD251	AL117191.6	14	CTGTGGAGGAAACATTGAAG	TCACACTCAAAGACTCCTTTC	IF	60	60	1995	172	288
L1AD252	AL590370.2	6	GTGAAGGGCACTGGTTATTA	TAATGAAATCGGACCTGTCT	FP	60	60	6498	408	202
L1AD253	AL163613.2	14	TTGCCTAGCTTTTCTACCAG	TTCAAGCTACCTTCTCAAGC	IF	60	60	1369	726	180
L1AD254	AL118557.5	14	ACCTTGACATTCCTGCAA	AATCCACCTGCAGACATTAC	FP	60	60	1000	143	514
L1AD255	AL117693.5	14	TCATTGTTCTATCCATGCCTTTT	GTAGGTTTGGGGCTGGAAAT	IF	55	60	961	197	228
L1AD256	AL161804.4	14	INSERTED IN REPEATS		R					
L1AD257	AL359545.12	10	INSERTED IN REPEATS		R					
L1AD258	AL358293.4	14	GGTTCAATTGAGCGTTACTT	TGCTGATATAGCACCTAGCA	FP	60	60	6800	735	300
L1AD259	AL158111.5	14	INSERTED IN REPEATS		R					
L1AD260	AL133238.3	14	GGTGGATGTATCCATTGTTT	TTTATGCATGCAAGAAATGA	FP	55	55	627	436	464
L1AD261	AL049838.3	14	CTATGGACCCATCTGACTGT	AGTTATTAAACCGGCCACTA	FP	60	60	6269	222	245

(Table cont.)

L1AD262	AC006568.7	4	ACACGGAGACACTTCAAATC	ACCCGTTATTGTGTTCAGAC	FP	60	60	6424	363	407
L1AD263	AL355481.12	13	GGCTACTTTGGCTTCTGTAA	ATTTGCTCAAACATTTCTGG	FP	55	55	5616	511	531
L1AD264	AL031681.16	20	GGGGAAGTTCCTCCTATATT	AAATGGTAGGTTGGTTTATCA	IF	60	60	1699	501	350
L1AD265	AL589693.3	6	ATAAATTTTCAGGCCTTTCC	GAACAAATTAGACACCATAAGGA	FP	60	60	6218	172	189
L1AD266	AL365508.19	6	INSERTED IN REPEATS		R					
L1AD267	AL445258.4	X	INSERTED IN REPEATS		R					
L1AD268	AL034425.9	20	GTTTAACCCAGCTGTCCAT	TCCTGTCTCATTTGCTTACC	FP	60	60	2022	361	395
L1AD269	AL136090.12	20	TGACATGGGAGCAATAATAGT	CAGGTGAAATGTATTGAAGGA	FP	55	55	1933	315	371
L1AD270	AL135936.11	20	INSERTED IN REPEATS		R					
L1AD271	AL390057.12	6	INSERTED IN REPEATS		NR					
L1AD272	AL161901.18	13	INSERTED IN REPEATS		R					
L1AD273	AC006947.2	17	GCCTGCTACATGTTCAGAT	CCATCCTTTCTGGAGTGAT	FP	60	60	6252	214	243
L1AD274	AL161938.6	20	INSERTED IN REPEATS		R					
L1AD275	AL157380.15	X	INSERTED IN REPEATS		R					
L1AD276	AL031679.1	20	ATTCTTCCTGCCACCTTATG	TTAATAGCTGAGCATCATGG	FP	60	60	993	492	372
L1AD277	AC006265.1	17	GTACAAACCATGGACCAGTT	ATGCAAGTATTTGGCATCTT	FP	55	60	6451	386	239
L1AD278	AL121757.7	UNK	INSERTED IN REPEATS		R					
L1AD279	AL157881.14	UNK	INSERTED IN REPEATS		R					
L1AD280	AC006131.1	UNK	INSERTED IN REPEATS		R					
L1AD281	AF036938.1	X	CAGAGTGAAGTGCTTGGTTT	CTTAATATTTGGGCCATGC	NR	60	55	1342	494	590
L1AD282	AL450303.10	1	NO RESULTS		NR					
L1AD283	AL358434.16	UNK	INSERTED IN REPEATS		R					
L1AD284	AL357141.8	6	NO RESULTS		NR					
L1AD285	AL359252.17	6	ATCCAATCACCATCATCAGT	ACCTGTGTTCCTATCTTTGC	FP	55	55	823	423	272
L1AD286	AL354937.12	9	TTTAACAACGCACACTTAGC	ATTAAGCAATGGCAGGAAT	FP	60	60	1385	337	444
L1AD287	AL356430.19	13	TTGAAATCAATAATGAGGGATA	AACATCAGTCAGCTAAAGCA	FP	55	55	518	277	256
L1AD288	AL121574.19	UNK	INSERTED IN REPEATS		R					
L1AD289	AL390039.10	UNK	INSERTED IN REPEATS		R					
L1AD290	AL158167.15	10	CCATGCCTCAACATCTCA	ACCTTCCTTATCTTCCCTTG	IF	60	60	750	175	237
L1AD291	AL157398.6	10	TGGAAAAATATCCCATAATGA	TTTCAGATGGTTTTTCAACA	FP	55	55	6277	180	311
L1AD292	AL136970.8	6	GGCAAATTGAGTCAAAGATG	AACTCATTCACAGTAGCAACAA	FP	60	60	6281	206	200

(Table cont.)

L1AD293	AL136117.12	6	TGGGAATCAGGAAATTTAAC	CCTATTTCTTGGGTTTTCTG	FP	60	60	2300	199	429
L1AD294	AL356286.8	X	INSERTED IN REPEATS		R					
L1AD295	AL158201.19	X	AAAGAAAGAAAACACCCACA	CTCACGTATTATTCCGATTTG	NP		60	2579	245	699
L1AD296	AL136441.16	13	AACCAAGGACTTACACATGC	ACTACCACTCATCCAGCAAA	FP	60	60	6518	461	261
L1AD297	AL357499.10	UNK	INSERTED IN REPEATS		R					
L1AD298	AL136455.6	1	TGCCACATCTGTTCAGTAAA	GAAATAGGCTCGTTTTCTCT	FP	60	60	1906	399	351
L1AD299	AL359502.14	13	TTAATGCAAGCAGAGTTTCC	TAAGAACCCATGGTCCAGTA	FP	60	55	6269	180	291
L1AD301	AL139334.10	6	AGTTGTCTGAGGAAACACCA	TACGCAGCATCAAGTAAAGA	FP	60	60	1823	700	288
L1AD303	AL139092.12	6	INSERTED IN REPEATS		R					
L1AD304	AC005358.1	17	ATCAGTGGTTCTTTGTCCTG	AGCAGTTCACAGTCCTTAGC	FP	55	55	1230	226	248
L1AD305	AC004768.1	5	GCCAGGAGATAATTTGTAGC	TACCTTGCCAGTAACCTTCT	FP	60	60	2726	386	330
L1AD306	AC004389.1	X	END OF CONTIG		EC					
L1AD307	AC004074.1	X	INSERTED IN REPEATS		R					
L1AD308	AC004523.1	UNK	INSERTED IN REPEATS		R					
L1AD309	AL138702.8	13	GCATTGCAGAAGAAAGCTA	TACCTCCAAGGCAAAACTTA	FP	60	60	1547	273	293
L1AD310	AL121946.20	6	CAACACACGTACAGGTATGC	TTAGCCTCTGTCTTTTGTGC	IF	60	55	6557	519	372
L1AD311	AL135932.7	11	TGACCTGTTCTGATGATTGA	CTTCTCAGGGTATCTGTCCA	FP	55	55	2281	271	327
L1AD312	AL136086.8	1	TTGGGGATAACTTTAACTGC	CCTTTTCATCCTCATGTTTT	IF	55	60	6284	228	209
L1AD313	AL137026.21	10	GCAGGAGAGAGTAAAGGGTTA	TGACAACCACTGCTATCAAG	FP	60	60	1382	86	165
L1AD314	AL121938.10	6	GGCTCAGGGAGATTTGATA	TCTGTTGTACTCTTTCAGGAACT	FP	60	55	3462	311	322
L1AD315	AL121933.15	6	GGTAACTAAAGCCATTGCAG	TATCTTTGGATGCTGCATAA	FP	55	55	2636	429	316
L1AD316	AL133547.16	9	INSERTED IN REPEATS		R					
L1AD317	AL157378.8	6	INSERTED IN REPEATS		R					
L1AD318	AL355871.5	1	TGTGGCTAATTCTGAGACCT	ACATGAGTTATCGTGGCATC	IF	60	60	631	176	175
L1AD319	AL157361.6	13	CCCAATGAACCTGTTGTAGT	GGATTTACATGCCACTTAGG	FP	55	60	392	188	241
L1AD320	AL157360.8	UNK	TCCAATGTTCTCTTAGAGGAGT	TCAACATGCAAAAGACTGAA	FP	60	55	489	114	248
L1AD321	AL139115.5	9	CTTGTCCATTTTCTCCACTG	CAACCCAGTAACTCCACTTC	FP	60	60	1193	80	200
L1AD322	AL049796.28	1	TTCTTCCTGGAAAATTGCTA	TTCCTATGAATCCAGTAGTGC	FP	55	60	6512	434	251
L1AD323	AL117345.21	6	GATGGCTTCAAATCCTTCTT	CACTTCAGATAGAACAAGAGCA	FP	60	55	3744	395	379
L1AD324	AL109920.15	6	TATCATTCCTTCAGGCCATA	GGTGAATGCTTTGGACTTTA	FP	55	60	1568	249	280
L1AD325	Z98950.1	X	TCGGCAGCACATATACTAAA	TCCATAGCCAAGTGAGTTTT	FP	60	55	1001	207	283

(Table cont.)

L1AD326	AL050309.4	X	INSERTED IN REPEATS		R					
L1AD327	AL030998.1	X	AAAACATATTTGGAGGAGCA	GTGACCTGGTGTTTTTGTCT	FP	55	55	6315	202	314
L1AD328	AL133353.6	10	TGCTAATAAAAGCACTCTGAAA	AAGATGGTGAATGTTGTAGGA	FP	55	60	2610	155	284
L1AD329	AL136169.6	UNK	INSERTED IN REPEATS		R					
L1AD330	AL133404.8	6	INSERTED IN REPEATS		R					
L1AD331	AL136363.4	X	ATTTCTTCTGCAGCTCTGAC	CATGATAACTTTGGTTTGTCAC	FP	60	60	6213	188	279
L1AD332	AL133247.1	2	TGACTGACCACTGTATGGAA	GTGGCTGTTTGGATTCTTTA	FP	60	60	1399	204	247
L1AD333	AL078604.10	6	INSERTED IN REPEATS		R					
L1AD334	AL021877.1	22	TTGACTTGTTTAGAAAGGGATT	GGATAAAGCTGAAAGCTCAA	FP	55	60	6322	233	215
L1AD335	Z70758.1	X	TCATCCAGCATTGAATCAG	TTGGTAGAAAGTGAAGTGGAG	FP	60	60	571	199	238
L1AD336	AL096706.10	UNK	INSERTED IN REPEATS		R					
L1AD337	AL049589.15	X	INSERTED IN REPEATS		R					
L1AD338	AL021069.1	1	AAGAATCCAATTTGCAACAG	TTTGATTCGGATTACACTGA	FP	60	60	6248	173	233
L1AD339	Z97181.1	X	GTTAAAATGCCAGGCTGAT	TGAGAAATGTGTTCTCCAAA	FP	55	55	1169	136	349
L1AD340	AL031117.1	X	INSERTED IN REPEATS		R					
L1AD341	AL034348.5	6	TGACTTCCATTTCAGGTACTC	CCACATTAGAGGTTTTCCAA	FP	55	60	4229	143	293
L1AD342	AL022399.2	1	TATGCATTTCCATGACTTGA	GTGGTAGGAGTAGGGGAAAG	FP	60	60	6795	342	708
L1AD343	AL033530.1	1	INSERTED IN REPEATS		R					
L1AD344	AL031313.1	X	INSERTED IN REPEATS		R					
L1AD345	AL023806.1	6	AGTACCAATGAAGTGCCATT	CAGGAGCATAAATAGGACCA	FP	60	60	1770	379	500
L1AD346	Z80232.1	X	CGGAAAATCCTCAGTCATC	ATGCCACAGCTTAAAAGTTC	FP	60	60	1065	261	309
L1AD347	Z84720.1	X	INSERTED IN REPEATS		R					
L1AD348	Z93018.1	X	NO RESULTS		NR					
L1AD349	Z99128.1	6	AGCACTCCTTTTATGAAGTCAACC	AGAGGAGAGAGTGGTTGATATTGG	FP	55	55	2851	1223	565
L1AD350	Z82170.1	UNK	GGCAGACCAAATGGATTAT	GATCCAAATATCAGACAAAATGT	FP	55	60	6342	288	184
L1AD351	Z95126.1	X	TGACATGCTTCCCTAAGTTT	TATAGAAAGTGAGGCCCAGA	FP	60	60	537	363	313
L1AD352	Z95325.2	X	CTTGCTGAATTAATCCCTTT	GGAAGAAATGATCCATAAGAAA	FP	55	55	3497	355	346
L1AD353	AL022308.1	X	CAAGGGGAAATCTCACAATA	GGACTTTGGGACTTACATCA	PARALOG	55	60	6238	174	263
L1AD354	AL023095.1	X	TCATCTTGCTCCCAAATATC	TCCTTAACACAGTCAAGTGAAC	FP	60	60	4839	170	338
L1AD355	Z98948.1	X	NO RESULTS		NR					
L1AD356	AC000111.1	7	TGTGGCTATGTGAGATGAGA	CCTTAATTTGAGGGGTTTTT	FP	55	55	4633	326	385

(Table cont.)

L1AD357	AP004241.2	11	CATAGGACGTTCAAGTGTGA	ATTGTCTATGGCTGCTTTCT	FP	60	55	765	387	593
L1AD358	AP002803.3	11	AGGTTTTGAGGTTTGCTGTA	TCCCAATAATCACTTTCCAC	FP	55	55	6274	205	264
L1AD359	AP002002.4	11	AAGGGCATATAAAACTGGTG	GCACCCATTAACTCATCATT	FP	55	60	6460	356	328
L1AD360	AP000764.4	11	CCATGCTTTCCACTCTTTAT	GCAGAAAAGGGTGTTCATA	FP	60	60	379	179	240
L1AD361	AP002784.3	11	GGAAAAATGACAGTCAGGAG	GCCTACCCAATGAATATCCT	HF	60	60	1031	149	258
L1AD362	AP003719.3	11	INSERTED IN REPEATS		R					
L1AD363	AP000811.4	11	CCATTACTTGAAGCAGAACC	CTGTGGGTCTCAGATCATTT	FP	55	55	6419	367	175
L1AD364	AP001977.4	11	TAAACTGGGGCTAGAAGTCA	CCAATTGAGAACCATCTTGT	FP	55	55	6335	383	344
L1AD365	AP002982.2	8	ACAGAGATTTCCTGGGCACT	TCAAACTGCATGCAAAATCC	FP	55	55	811	109	208
L1AD367	AP000789.4	11	CCAACAGGGATCAAAGGTTC	GCCACCTTGAGTTGGTGAAG	FP	55	55	378	147	175
L1AD368	AP002006.5	11	TTTCTTTTCCTACTCTCCCTCTC	GAGAAATAAAGGCAATTGCTCAC	NP		55	4593	186	922
L1AD369	AP001485.4	11	AAAACATATAAGCGGCCAAC	CAGCACCTGTTATGGTTTGA	FP	60	55	2437	466	187
L1AD370	AP000462.2	11	TAAGAAGAGGGGAGGAGACT	GCCTCTATGAAGCAGGTATG	FP	55	60	793	178	237
L1AD371	AP001709.1	21	CTAAATTGCTCCATTCCTTG	ATCACTGTAGGGTGATCCAG	HF	55	55	2525	581	562
L1AD372	AP001678.1	21	CTTACGCCTCAATTATCTGG	TGCAATTGATCTTACAAGGA	FP	55	55	2325	280	269
L1AD373	AP001674.1	21	CAAATAGCCAGCACAATATG	TTGTCATTGGTCTTTTGTCA	FP	55	60	823	165	226
L1AD374	AP001669.1	21	INSERTED IN REPEATS		R					
L1AD375	AB009801.1	14	AATCCACCTGCAGACATTAC	AGAACATCCCCTATCCAAAC	FP	55	55	688	87	202
L1AD382	Z95325.2	X	INSERTED IN REPEATS		R					
L1AD383	AC090791.6	11	TGGTGGTCTCAGAGTAAACA	ACCCAAAACATCATTAGTGC	FP	60	60	1642	117	1026
L1AD384	AL136441.16	13	INSERTED IN REPEATS		R					
L1AD385	AP003123.2	11	GCACAGGTTTATCTCCTTGA	ATTGAAGACCTGCAATTTGT	FP	55	55	6379	284	287
L1AD386	AC114975.2	5	INSERTED IN REPEATS		R					
L1ADY8	AC010970.3	Y	TCACACGTATCCCTTTGCAG	TTTTCTGTGAAACATCTTGGAGA	FP	55	55	1813	115	204

* Indicates L1 preTa element identified by Ovchinnikov 2002

1. Chromosomal location was determined from Accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples. L1 elements with unknown locations are denoted UNK.
2. Elements at the end of sequencing contigs are denoted (EC), those residing in other repeats (R), those having paralogs (PARALOG), and elements with inconclusive PCR results (NR). Elements represented here are classified according to allele

frequency as: high frequency (HF), intermediate (IF), No pre-integration site in primate samples tested (NP), or as fixed present (FP) insertions. Fixed present: every individual tested had the LINE element in both chromosomes. Intermediate frequency insertion polymorphism: the element present in more than 30% of alleles tested and no more than 70% of the alleles. High frequency insertion polymorphism: the element is present in more than 70% but not all alleles tested. Indeterminable data is denoted (-).

3. Amplification of each locus required 2:30 min @ 94°C initial denaturing, and 32 cycles for 1 min 94°C, 1 min Annealing Temperature (A.T.), and 1 min elongation at 72°C. A final extension time of 10 min at 72°C was also used.
4. PCR product sizes: Empty product size is calculated computationally by removing the L1 preTa elements and 1 direct repeat from identified filled site. Subfamily specific product size is calculated from internal subfamily specific primer located in the 3' UTR to the proximal 3' primer. In cases where target site duplication sequence were not found flanking the element PCR product sizes may vary from those reported.

Table 4 - Autosomal preTa L1 allele frequency and heterozyosity.

	African American					Asian/Alaskan Native[d]					German Caucasian					Egyptian					
	Genotypes					Genotypes					Genotypes					Genotypes					
Elements	+/+	+/-	-/-	**f[c]**	**H[a]**	+/+	+/-	-/-	**f[c]**	**H[a]**	+/+	+/-	-/-	**f[c]**	**H[a]**	+/+	+/-	-/-	**f[c]**	**H[a]**	**AH[b]**
L1AD10	0	5	14	0.13	0.23	0	8	12	0.20	0.33	3	7	7	0.38	0.49	3	7	10	0.33	0.45	0.37
L1AD14	9	10	1	0.70	0.43	4	8	7	0.42	0.50	16	4	0	0.90	0.18	17	2	1	0.90	0.18	0.33
L1AD19	13	7	0	0.83	0.30	15	2	0	0.94	0.11	14	6	0	0.85	0.26	14	6	0	0.85	0.26	0.23
L1AD20	18	2	0	0.95	0.10	19	1	0	0.98	0.05	16	0	0	1.00	0.00	19	0	0	1.00	0.00	0.04
L1AD75	0	5	15	0.13	0.22	0	1	18	0.03	0.05	1	9	9	0.29	0.42	0	9	11	0.23	0.36	0.26
L1AD77	1	5	11	0.21	0.34	0	1	19	0.03	0.05	0	3	13	0.09	0.18	0	6	12	0.17	0.29	0.21
L1AD82	19	1	0	0.98	0.05	17	0	0	1.00	0.00	20	0	0	1.00	0.00	19	1	0	0.98	0.05	0.03
L1AD96	13	5	2	0.78	0.36	15	1	0	0.97	0.06	5	10	5	0.50	0.51	11	7	2	0.73	0.41	0.34
L1AD100	19	0	0	1.00	0.00	19	0	1	0.95	0.10	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.02
L1AD101	16	4	0	0.90	0.18	10	5	0	0.83	0.29	13	6	1	0.80	0.33	11	9	2	0.70	0.43	0.31
L1AD102	14	0	0	1.00	0.00	14	1	0	0.97	0.07	12	1	2	0.83	0.29	0	4	16	0.10	0.18	0.13
L1AD125	12	7	1	0.78	0.36	14	6	0	0.85	0.26	20	0	0	1.00	0.00	19	1	0	0.98	0.05	0.17
L1AD135	19	1	0	0.98	0.05	20	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.01
L1AD160	11	5	4	0.68	0.45	5	11	4	0.53	0.51	4	12	1	0.59	0.50	4	8	3	0.53	0.51	0.49
L1AD176	7	3	2	0.71	0.43	2	9	5	0.41	0.50	0	1	15	0.03	0.06	1	0	11	0.08	0.16	0.29
L1AD186	4	7	8	0.39	0.49	14	5	1	0.83	0.30	5	10	2	0.59	0.50	4	11	5	0.48	0.51	0.45

(Table cont.)

L1AD189	14	5	0	0.87	0.23	19	0	0	1.00	0.00	20	0	0	1.00	0.00	19	1	0	0.98	0.05	0.07
L1AD208	14	6	0	0.85	0.26	19	0	0	1.00	0.00	14	0	0	1.00	0.00	14	0	0	1.00	0.00	0.07
L1AD213	7	9	3	0.61	0.49	2	12	5	0.42	0.50	2	2	5	0.33	0.47	8	5	7	0.53	0.51	0.49
L1AD219	3	14	3	0.50	0.51	0	10	10	0.25	0.38	1	5	14	0.18	0.30	2	11	7	0.38	0.48	0.42
L1AD230	14	6	0	0.85	0.26	19	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.07
L1AD232	13	7	0	0.83	0.30	8	7	3	0.64	0.47	12	2	0	0.93	0.14	13	4	1	0.83	0.29	0.30
L1AD240	13	3	0	0.91	0.18	20	0	0	1.00	0.00	13	0	0	1.00	0.00	20	0	0	1.00	0.00	0.04
L1AD251	3	9	7	0.39	0.49	10	8	2	0.70	0.43	14	4	0	0.89	0.20	8	11	1	0.68	0.45	0.39
L1AD253	11	6	3	0.70	0.43	0	14	5	0.37	0.48	4	8	7	0.42	0.50	0	6	14	0.15	0.26	0.42
L1AD255	1	8	10	0.26	0.40	1	9	10	0.28	0.41	6	7	7	0.48	0.51	3	14	3	0.50	0.51	0.46
L1AD264	4	10	6	0.45	0.51	2	9	8	0.34	0.46	2	7	7	0.34	0.47	3	11	6	0.43	0.50	0.48
L1AD290	7	12	1	0.65	0.47	4	8	7	0.42	0.50	3	13	0	0.59	0.50	6	9	5	0.53	0.51	0.49
L1AD310	5	6	7	0.44	0.51	0	5	15	0.13	0.22	5	2	5	0.50	0.52	6	5	7	0.47	0.51	0.44
L1AD312	0	4	16	0.10	0.18	11	6	2	0.74	0.40	2	9	5	0.41	0.50	2	7	9	0.31	0.44	0.38
L1AD318	4	12	4	0.50	0.51	2	12	6	0.40	0.49	4	8	8	0.40	0.49	3	11	6	0.43	0.50	0.50
L1AD361	17	3	0	0.93	0.14	19	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.04
L1AD371	15	5	0	0.88	0.22	18	2	0	0.95	0.10	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.08
L1AD264	4	10	6	0.45	0.51	2	9	8	0.34	0.46	2	7	7	0.34	0.47	3	11	6	0.43	0.50	0.48
L1AD290	7	12	1	0.65	0.47	4	8	7	0.42	0.50	3	13	0	0.59	0.50	6	9	5	0.53	0.51	0.49
L1AD310	5	6	7	0.44	0.51	0	5	15	0.13	0.22	5	2	5	0.50	0.52	6	5	7	0.47	0.51	0.44
L1AD312	0	4	16	0.10	0.18	11	6	2	0.74	0.40	2	9	5	0.41	0.50	2	7	9	0.31	0.44	0.38
L1AD318	4	12	4	0.50	0.51	2	12	6	0.40	0.49	4	8	8	0.40	0.49	3	11	6	0.43	0.50	0.50
L1AD361	17	3	0	0.93	0.14	19	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.04
L1AD371	15	5	0	0.88	0.22	18	2	0	0.95	0.10	20	0	0	1.00	0.00	20	0	0	1.00	0.00	0.08

a. This is unbiased heterozygosity. H=(2*sample size * (1-sum of homozygotes))/(2*sample size – 1)
b. Average heterozygosity is the average heterozygosity for all populations.
c. f represents the frequency of the element.
d. Asian and Alaskan Native samples were used interchangeably as a geographically unique human population

VITA

Jeremy Shawn Myers is the son of Joseph E. Myers and Joanne M. Lindner. He graduated with a Bachelor of Science degree in biology from Bucknell University in Lewisburg, Pennsylvania, in May of 1999. He completed the first two years of his doctoral work at Louisiana State University Health Sciences Center in New Orleans, Louisiana. He transferred to Louisiana State University in Baton Rouge, Louisiana, in the spring of 2001. Mr. Myers will graduate with the degree of Doctor of Philosophy in biochemistry from Louisiana State University in May, 2003.